SUPPLEMENT TO "DOUBLE-SLICING ASSISTED SUFFICIENT DIMENSION REDUCTION FOR HIGH DIMENSIONAL CENSORED DATA"

By Shanshan Ding, Wei Qian and Lan Wang

I. Lemmas, theorems and proofs. We first provide proofs of the propositions in Supplement I.1 and briefly verify (3.1). Some useful lemmas are given in Supplement I.2 as preliminary results for Supplements I.3– I.5. Supplement I.3 provides proofs for Theorems 4.2 and 4.4, and Supplement I.4 provides the proofs of Theorems 4.1 and 4.3. Supplement I.5 gives additional theoretical results on the local KM estimator.

I.1. *Proofs of Propositions.*

Proof of Proposition A.1. Let $\mathcal{A}$ be any index set in $\{1, \cdots, p\}$, and define $S_{\mathcal{A}}$ to be the $p \times |\mathcal{A}|$ submatrix extracted from identity matrix I_p according to column index $\mathcal{A}$. Then, following Yin & Hilafu (2015), similar to the central subspace, we can define the central variable selection subspace $\mathcal{S}_0$ to be the intersection of all subspaces $\mathrm{Span}(S_{\mathcal{A}})$ with $\mathcal{A}$ satisfying $T \perp\!\!\!\perp \mathbf{X}_{\mathcal{A}^c} \,|\, \mathbf{X}_{\mathcal{A}}$. By the existence of $\mathcal{S}_{T|\mathbf{X}}$, $\mathcal{S}_0$ also exists (and is unique). Consequently, there is a unique index set $\mathcal{A}_0$ in $\{1, \cdots, p\}$ such that $\mathcal{S}_0 = \mathrm{Span}(S_{\mathcal{A}_0})$ and $T \perp\!\!\!\perp \mathbf{X}_{\mathcal{A}_0^c} \,|\, \mathbf{X}_{\mathcal{A}_0}$. By definition, $\mathcal{S}_{T|\mathbf{X}} \subseteq \mathcal{S}_0$; consequently, for any basis matrix Γ_0 of $\mathcal{S}_{T|\mathbf{X}}$, $\boldsymbol{\gamma}_{0j} = \mathbf{0}$ for any $j \in \mathcal{A}_0^c$,where $\boldsymbol{\gamma}_{0j}$ is the jth row of Γ_0. In addition, given any $j \in \mathcal{A}_0$, $\|\boldsymbol{\gamma}_{0j}\| > 0$; indeed, if $\|\boldsymbol{\gamma}_{0j}\| = 0$, then $T \perp\!\!\!\perp \mathbf{X} \,|\, \mathbf{X}_{\mathcal{A}_0 \setminus \{j\}}$, which contradicts the definition of $\mathcal{S}_0$. This completes the proof of Proposition A.2. □

Proof of Proposition A.2. First, we show that there exists a minimizer Γ_0 that satisfies $\mathrm{Span}(\Gamma_0) = \mathcal{S}_{T|\mathbf{X}}$. Indeed, since $\dim(\Sigma^{-1}U_c) = \dim(\mathcal{S}_{Y|\mathbf{X}}) = d$, there must exist full-rank matrices $\tilde{\Gamma} \in \mathbb{R}^{p\times d}$ and $\tilde{\Phi}^{d\times b}$ such that $\Sigma^{-1}U_c = \tilde{\Gamma}\tilde{\Phi}$. Then with $Q = (\tilde{\Phi}\tilde{\Phi}^T)^{\frac{1}{2}}$, we can construct $(\Gamma_0, \Phi_0) = (\tilde{\Gamma}Q, Q^{-1}\tilde{\Phi})$. Apparently, $\mathrm{Span}(\Gamma_0) = \mathrm{Span}(\tilde{\Gamma}) = \mathrm{Span}(\mathcal{S}_{T|X})$, and (Γ_0, Φ_0) forms a minimizer of (A.1) since $F(\Gamma, \Phi)$ is always non-negative. Next, we show that the minimizer is unique up to an orthogonal transformation. Suppose $(\bar{\Gamma}_0, \bar{\Phi}_0)$ is another minimizer of (A.1). Then by the constraint on Φ, there exists $\bar{Q} \in \mathbb{R}^{d\times d}$ such that $\bar{\Phi}_0 = \bar{Q}\Phi_0$ and $\bar{Q}$ is orthogonal. Therefore,

$(\Gamma_0\bar{Q}^{-1}, \bar{\Phi}_0)$ is also a minimizer of (A.1). By convexity of $F(\Gamma, \bar{\Phi}_0)$ with respect to Γ, we have $\bar{\Gamma}_0 = \Gamma_0\bar{Q}^{-1}$ and consequently, $\mathrm{Span}(\bar{\Gamma}_0) = \mathrm{Span}(\mathcal{S}_{T|X})$. This completes the proof of Proposition A.2. $\square$

Proof of Proposition A.3. By law of iterated expectation, $T \perp\!\!\!\perp C \,|\, \mathbf{X}$ and $T \perp\!\!\!\perp \mathbf{X} \,|\, \mathbf{X}_{\mathcal{A}_0}$, given any $t, c > 0$,

$$\begin{aligned} P(T \le t,\, C \le c \,|\, \mathbf{X}_{\mathcal{A}_0}) &= \mathrm{E}(P(T \le t,\, C \le c \,|\, \mathbf{X}) \,|\, \mathbf{X}_{\mathcal{A}_0}) \\ &= \mathrm{E}(P(T \le t \,|\, \mathbf{X}) P(C \le c \,|\, \mathbf{X}) \,|\, \mathbf{X}_{\mathcal{A}_0}) \\ &= \mathrm{E}(P(T \le t \,|\, \mathbf{X}_{\mathcal{A}_0}) P(C \le c \,|\, \mathbf{X}) \,|\, \mathbf{X}_{\mathcal{A}_0}) \\ &= P(T \le t \,|\, \mathbf{X}_{\mathcal{A}_0}) P(C \le c \,|\, \mathbf{X}_{\mathcal{A}_0}). \end{aligned}$$

Therefore, $T \perp\!\!\!\perp C \,|\, \mathbf{X}_{\mathcal{A}_0}$. Using similar arguments as before, we obtain that $T \perp\!\!\!\perp C \,|\, \Gamma_0^T\mathbf{X}$. This completes the proof of Proposition A.3. $\square$

Proof of Proposition A.4. To show the first inequality in (A.2), let Γ_{01} be the $p \times d$ matrix consisting of left singular vectors of $\tilde{U}_c$ corresponding to nonzero singular values. Then with SVD of $\hat{\Gamma}_{01}^T\Gamma_{01}$, we can obtain LVR^T, where L and R are orthogonal matrices and $V = \mathrm{diag}(\cos\theta_1, \cdots, \theta_d)$, and $\theta_1, \cdots, \theta_d$ are the principal angles between $\mathcal{S}_{\hat{\Xi}}$ and $\mathcal{S}_{T|\mathbf{X}}$. Then, defining $Q = RL^T$ and $\Gamma_0 = \Gamma_{01}Q$, we have

$$\begin{aligned} &\|\hat{\Gamma}_{01} - \Gamma_0\|_F^2 = \mathrm{tr}\big[(\hat{\Gamma}_{01} - \Gamma_{01}Q)^T(\hat{\Gamma}_{01} - \Gamma_{01}Q)\big] = 2d - 2\mathrm{tr}(\hat{\Gamma}_{01}^T\Gamma_{01}Q) \\ =&2d - 2\sum_{j=1}^{d}\cos\theta_j \le 2\sum_{j=1}^{d}\sin^2\theta_j = 2\|\sin\Theta(\mathcal{S}_{\hat{\Xi}}, \mathcal{S}_{T|\mathbf{X}})\|_F^2. \end{aligned}$$

The second inequality in (A.2) then follows immediately from Wedin's theorem. $\square$

Proof of Proposition A.5. First, since both $\mathcal{S}_{T|\mathbf{X}}$ and $\mathcal{S}_{C|\mathbf{X}}$ are contained in $\mathcal{S}_{(T,C)|\mathbf{X}}$, it is apparent that the former condition implies the latter. Next, we want to show the latter conditions imply the former one. Let Γ_0 and Γ_c be the basis of $\mathcal{S}_{T|\mathbf{X}}$ and $\mathcal{S}_{C|\mathbf{X}}$, respectively. Then we have by the latter conditions that $\Gamma_0 \in \mathcal{S}_{(Y,\delta)|\mathbf{X}}$ and $\Gamma_c \in \mathcal{S}_{(Y,\delta)|\mathbf{X}}$, which implies that $\mathrm{Span}(\Gamma_0, \Gamma_c) \subseteq \mathcal{S}_{(Y,\delta)|\mathbf{X}}$. In addition, we note that $\mathrm{Span}(\Gamma_0, \Gamma_c) \equiv \mathcal{S}_{(T,C)|\mathbf{X}}$ and consequently, $\mathcal{S}_{(T,C)|\mathbf{X}} \subseteq \mathcal{S}_{(Y,\delta)|\mathbf{X}}$. Indeed, it is easy to see by definition that $\mathrm{Span}(\Gamma_0, \Gamma_c) \subseteq \mathcal{S}_{(T,C)|\mathbf{X}}$. Also, note by law of iterated expectation and

$T \perp\!\!\!\perp C \mid \mathbf{X}$ that given any $t, c > 0$

$$\begin{aligned}P(T \le t,\, C \le c \mid \Gamma_0^T\mathbf{X},\, \Gamma_c^T\mathbf{X}) &= \mathrm{E}(P(T \le t,\, C \le c \mid \mathbf{X}) \mid \Gamma_0^T\mathbf{X},\, \Gamma_c^T\mathbf{X}) \\ &= \mathrm{E}(P(T \le t \mid \mathbf{X})P(C \le c \mid \mathbf{X}) \mid \Gamma_0^T\mathbf{X},\, \Gamma_c^T\mathbf{X}) \\ &= \mathrm{E}(P(T \le t \mid \Gamma_0^T\mathbf{X})P(C \le c \mid \Gamma_c^T\mathbf{X}) \mid \Gamma_0^T\mathbf{X},\, \Gamma_c^T\mathbf{X}) \\ &= P(T \le t \mid \Gamma_0^T\mathbf{X})P(C \le c \mid \Gamma_c^T\mathbf{X}) \\ &= P(T \le t \mid \mathbf{X})P(C \le c \mid \mathbf{X}) \\ &= P(T \le t,\, C \le c \mid \mathbf{X}),\end{aligned}$$

which implies that $\mathrm{Span}(\Gamma_0, \Gamma_c) \supseteq \mathcal{S}_{(T,C)|\mathbf{X}}$. This completes the proof of Proposition A.5. □

PROOF OF PROPOSITION A.6. By the same arguments as Proposition A.2, it is not hard to show that Γ_1 is unique up to an orthogonal transformation, which immediately implies that $\|\Gamma_{1,\mathcal{A}}\|_{2,1}$ is also unique given $\mathcal{A}$. □

The following is a brief derivation of (3.1).

$$\begin{aligned}\mathrm{E}\Big\{\frac{\delta(\mathbf{X}-\boldsymbol{\mu})}{S(Y \mid \mathbf{X})}\,\Big|\, Y \in \mathcal{H}_y\Big\} &= \mathrm{E}\Big\{\mathrm{E}\Big\{\frac{\delta(\mathbf{X}-\boldsymbol{\mu})I(Y \in \mathcal{H}_y)}{S(Y \mid \mathbf{X})}\,\Big|\, T, \mathbf{X}\Big\}\Big\}/p_y \\ &= \mathrm{E}\Big\{\mathrm{E}\Big\{\frac{I(T \le C)(\mathbf{X}-\boldsymbol{\mu})I(T \in \mathcal{H}_y)}{S(T \mid \mathbf{X})}\,\Big|\, T, \mathbf{X}\Big\}/p_y \\ &= \mathrm{E}\Big\{(\mathbf{X}-\boldsymbol{\mu})I(T \in \mathcal{H}_y)\Big\}/p_y = \frac{\tilde{p}_y}{p_y}\mathrm{E}(\mathbf{X}-\boldsymbol{\mu} \mid T \in \mathcal{H}_y).\end{aligned}$$

I.2. *Lemmas.*

LEMMA I.1. *Given $\epsilon_n \to 0$ and $n\epsilon_n^2 \to \infty$, for slice $\mathcal{H}_y$, with large enough n,*

$$P(|\hat{p}_y - p_y| > \epsilon_n) \le 2\exp(-n\epsilon_n^2).$$

In addition, under Condition (A3),

$$P(|\bar{X}_j - \mu_j| > \epsilon_n) \le 2\exp\Big(-\frac{n\epsilon_n^2}{2C}\Big),$$

for large enough n and some positive constant C.

Lemma I.1 follows directly from Bernstein inequality and Hoeffding's inequality. We omit its proof.

LEMMA I.2. *Suppose Conditions (A2), (A3) and (C4) are satisfied. Given $\epsilon_n \to 0$, $n\epsilon_n^2/\bar{b} \to \infty$ and $1 \le j \le p$, for slice $\mathcal{H}_y$ with large enough n,*

$$P(|\bar{X}_{yj} - \mu_{yj}| > \epsilon_n) \le \frac{32v^2}{\epsilon_n^2}\exp\Big(-\frac{np_y\epsilon_n^2}{v^2}\Big), \tag{I.1}$$

where $\bar{X}_{yj}$ is the jth element of $\bar{\mathbf{X}}_y$, μ_{yj} is the jth element of $\boldsymbol{\mu}_y := \mathrm{E}(\mathbf{X} \,|\, Y \in \mathcal{H}_y)$, and v is some positive constant.

Proof of Lemma I.2. Note that by Condition (A3), we know that there exist positive constants v and c_0 such that $\mathrm{E}(|X_j - \mu_{yj}|^k \,|\, Y \in \mathcal{H}_y) \leq k! v^2 c_0^{k-2}/2$. Then,

$$
\begin{aligned}
& P(|\bar{X}_{yj} - \mu_{yj}| > \epsilon_n) \\
\leq & P\Big(\frac{N_y}{n} \leq \frac{p_y}{2}\Big) + P\Big(|\bar{X}_{yj} - \mu_{yj}| > \epsilon_n, \, \frac{N_y}{n} > \frac{p_y}{2}\Big) \\
\leq & \exp\Big(-\frac{3np_y}{28}\Big) + \sum_{j=\lceil np_y/2\rceil}^{n} P(|\bar{X}_{yj} - \mu_{yj}| > \epsilon_n, \, N_y = j) \qquad \text{(I.2)} \\
\leq & \exp\Big(-\frac{3np_y}{28}\Big) + \sum_{j=\lceil np_y/2\rceil}^{n} 2\exp\Big(-\frac{j\epsilon_n^2}{2(v^2 + c_0\epsilon_n)}\Big) \qquad \text{(I.3)} \\
\leq & \exp\Big(-\frac{3np_y}{28}\Big) + 2\exp\Big(-\frac{np_y\epsilon_n^2}{4(v^2 + c_0\epsilon_n)}\Big) \Big/ \Big(1 - \exp\big(-\frac{\epsilon_n^2}{2(v^2 + c_0\epsilon_n)}\big)\Big) \\
\leq & \exp\Big(-\frac{3np_y}{28}\Big) + \frac{16v^2}{\epsilon_n^2}\exp\Big(-\frac{np_y\epsilon_n^2}{v^2}\Big),
\end{aligned}
$$

where (I.2) and (I.3) hold by the extended Bernstein inequalities (see, e.g., Lemma 1 and Lemma 2 in Qian & Yang 2016). In together with Conditions (A2) and (C4), we obtain (I.1). □

The following Lemma I.3 is a direct consequence of Lemma 19.15 in van der Vaart (2000). Suppose $X_1, \cdots X_n$ are i.i.d. random variables with probability measure P, and P_n is the corresponding empirical measure. Given a measurable function g, define $Pg = \mathrm{E}[g(X_1)]$ and and $P_n g = \frac{1}{n}\sum_{i=1}^n g(X_i)$.

Lemma I.3. *There exists a universal constant $K > 0$ such that for any VC-class $\mathcal{F}$ of functions, for every $0 < \epsilon < 1$,*

$$
\mathcal{N}(\epsilon\|F\|_{P_n,1}, \mathcal{F}, L_1(P_n)) \leq K V_1 (16e)^{V_1} \epsilon^{-(V_1 - 1)}, \qquad \text{(I.4)}
$$

where V_1 is VC-index of $\mathcal{F}$, F is the envelope function, $\|\cdot\|_{P_n,1}$ is the L_1 norm under empirical measure, and $\mathcal{N}(\epsilon, \mathcal{F}, L_1(P_n))$ is the ϵ-covering number of $\mathcal{F}$ under the empirical L_1 norm. In particular, if $\mathcal{F} = \{1_{(t,\infty)} : t \in [0, T_0)\}$, then

$$
\log \mathcal{N}(\epsilon, \mathcal{F}, L_1(P_n)) \leq a_1 \log(\frac{1}{\epsilon}) + a_2, \qquad \text{(I.5)}
$$

where a_1 and a_2 are some positive constants.

LEMMA I.4. *Let $\mathcal{F}$ be a class of functions with ϵ-covering number $N_\epsilon := \mathcal{N}(\epsilon, \mathcal{F}, L_1(P_n))$. Assume $\max_{f\in\mathcal{F}}\|f\|_\infty \le 1$. Then for every $\epsilon \ge 2\sqrt{\frac{\log N_\epsilon}{n}}$,*

$$P\big(\sup_{f\in\mathcal{F}}|P_n f - Pf| > 8\epsilon\big) \le 8\exp\big(-\frac{n\epsilon^2}{4}\big).$$

PROOF OF LEMMA I.4. Let $\mathcal{G}$ be an ϵ-covering of $\mathcal{F}$ under $L_1(P_n)$. Let $\{W_i\}_{i=1}^n$ be a Rademacher sequence and define $\tilde{P}_n f = \frac{1}{n}\sum_{i=1}^n W_i f(X_i)$. Then

$$\sup_{f\in\mathcal{F}}|\tilde{P}_n f| \le \max_{f\in\mathcal{G}}|\tilde{P}_n f| + \epsilon,$$

which, in together with Lemma 4.1.1 in van de Geer (2006), implies that

$$P\big(\sup_{f\in\mathcal{F}}|\tilde{P}_n f| > 2\epsilon\big) \le P\big(\max_{f\in\mathcal{G}}|\tilde{P}_n f| > \epsilon\big) \le 2\exp\big(-\frac{n\epsilon^2}{4}\big). \tag{I.6}$$

In addition, by Chebyshev's inequality, for every $f \in \mathcal{F}$,

$$P(|P_n f - Pf| > 4\epsilon) \le \frac{\mathrm{Var}(f(X_1))}{16n\epsilon^2} \le \frac{1}{64\log N_\epsilon} \le \frac{1}{2}. \tag{I.7}$$

Then by (I.6), (I.7) and Corollary 3.2.2 in van de Geer (2006),

$$P\big(\sup_{f\in\mathcal{F}}|P_n f - Pf| > 8\epsilon\big) \le 4P\big(\sup_{f\in\mathcal{F}}|\tilde{P}_n f| > 2\epsilon\big) \le 8\exp\big(-\frac{n\epsilon^2}{4}\big),$$

which completes the proof of Lemma I.4. □

I.3. *Proofs of Theorems 4.2 and 4.4.* The proof of Theorem 4.2 uses key results given in the following Theorem I.1.

THEOREM I.1. *Suppose conditions in Theorem 4.1 and Conditions (C1)-(C6) hold and take $h_n \asymp (\frac{\log\bar{p}_n}{n})^{\frac{1}{2+d_1}}$. Assume that for every $j \in \mathcal{A}_0$, $\lambda w_j \le 2c_0\xi_n$ and for every $j \in \mathcal{A}_0^c$, $\lambda w_j \ge 2c_0\xi_n$, where $\xi_n = b^{1/2}(\log\bar{p}_n/n)^{\frac{1}{2+d_1}} + (\bar{b}q_1^{1/2} + q_1 + b^{1/2})(b\log\bar{p}_n/n)^{1/2}$ and c_0 is a constant given in (I.16). Then there exist constants $\tilde{C}_1, \tilde{C}_2 > 0$ such that for large enough n, with probability greater than $1 - 1/\bar{p}_n$,*

$$\|\mathrm{P}_{\hat{\Gamma}_0} - \mathrm{P}_{\Gamma_0}\|_F \le \tilde{C}_1(bq)^{1/2}\big(\frac{\log\bar{p}_n}{n}\big)^{\frac{1}{2+d_1}} + \tilde{C}_2(\bar{b}q_1^{1/2} + q_1 + b^{1/2})\big(\frac{bq\log\bar{p}_n}{n}\big)^{1/2}.$$

PROOF OF THEOREM I.1. Since $F_n(\hat{\Gamma}_0, \hat{\Phi}_0) \le F_n(\Gamma_0, \Phi_0)$, we have

$$\begin{aligned} &\tilde{G}_1 + \tilde{G}_2 + \tilde{G}_3 + \lambda \sum_{j=1}^{p} w_j \|\hat{\boldsymbol{\gamma}}_{0j}\|_2 \\ :=& - \mathrm{tr}[(\hat{U}_c - \Sigma\Gamma_0\Phi_0)^T \hat{\Delta}_0] + \mathrm{tr}[\Phi_0^T \Gamma_0^T (\Sigma_n - \Sigma)\hat{\Delta}_0] \\ &+ \frac{1}{2}\mathrm{tr}(\hat{\Delta}_0^T \Sigma_n \hat{\Delta}_0) + \lambda \sum_{j=1}^{p} w_j \|\hat{\boldsymbol{\gamma}}_{0j}\|_2 \le \lambda \sum_{j=1}^{p} w_j \|\boldsymbol{\gamma}_{0j}\|_2, \end{aligned} \tag{I.8}$$

which follows straightforwardly from arguments similar to (I.34), where $\hat{\Delta}_0 := \hat{\Gamma}_0\hat{\Phi}_0 - \Gamma_0\Phi_0 = (\hat{\boldsymbol{\delta}}_{01}, \cdots, \hat{\boldsymbol{\delta}}_{0p})^T$. Define $\tilde{\mathbf{m}}_y = N_y^{-1}\sum_{i\in J_y} \delta_i(\mathbf{X}_i - \boldsymbol{\mu})/S_{\mathbf{X}_i}(Y_i)$, $\hat{\mathbf{m}}_{2y} = \hat{g}_y \hat{\mathbf{m}}_y$ and $\mathbf{m}_{2y} = g_y \mathbf{m}_y$ for $1 \le y \le b$. Given $1 \le j \le p$, let $\hat{m}_{yj}$, $\hat{m}_{2yj}$ and $\tilde{m}_{yj}$ be the jth element of $\hat{\mathbf{m}}_y$, $\hat{\mathbf{m}}_{2y}$ and $\tilde{\mathbf{m}}_y$, respectively. Then

$$\begin{aligned} &|\hat{m}_{2yj} - m_{2yj}| \\ =&\Big|\frac{\hat{p}_y - p_y}{\sqrt{\hat{p}_y} + \sqrt{p_y}}(\hat{m}_{yj} - m_{yj}) + \sqrt{p_y}(\hat{m}_{yj} - m_{yj}) + \frac{\hat{p}_y - p_y}{\sqrt{\hat{p}_y} + \sqrt{p_y}} m_{yj}\Big| \\ \le& c_s^{-1/2} b^{1/2} |(\hat{p}_y - p_y)(\hat{m}_{yj} - m_{yj})| + |\hat{m}_{yj} - m_{yj}| + 2c_s^{-1/2} K b^{1/2} |\hat{p}_y - p_y|. \end{aligned} \tag{I.9}$$

To provide upper bound for the display above, first note that if $|\hat{S}(Y_i \,|\, \mathbf{X}_i) - S(Y_i \,|\, \mathbf{X}_i)| \le \tau_1/2 := \tau_0^2/2$ for every $1 \le i \le n$, then

$$\begin{aligned} &|\hat{m}_{yj} - m_{yj}| = |\hat{m}_{yj} - \tilde{m}_{yj} + \tilde{m}_{yj} - m_{yj}| \\ \le& \Big|\frac{1}{N_y}\sum_{i\in J_y} \delta_i (X_{ij} - \mu_j)\Big(\frac{1}{S(Y_i \,|\, \mathbf{X}_i)} - \frac{1}{\hat{S}(Y_i \,|\, \mathbf{X}_i)}\Big)\Big| \\ &+ \Big|\frac{1}{N_y}\sum_{i\in J_y} \frac{\delta_i(\bar{X}_j - \mu_j)}{\hat{S}(Y_i \,|\, X_i)}\Big| + |\tilde{m}_{yj} - m_{yj}| \\ \le& 4K\tau_1^{-2} |S(Y_i \,|\, \mathbf{X}_i) - \hat{S}(Y_i \,|\, \mathbf{X}_i)| + 2\tau_1^{-1} |\bar{X}_j - \mu_j| + |\tilde{m}_{yj} - m_{yj}|. \end{aligned} \tag{I.10}$$

Taking $h_n = (\frac{\log \bar{p}_n}{n})^{\frac{1}{2+d_1}}$ and $\epsilon_n = \alpha\sqrt{\frac{\log \bar{p}_n}{n h_n^{d_1}}}$ for some constant $\alpha > 0$, we have by results in Lemma I.6, (I.41) and Wedin's theorem (e.g., Yu et al. 2015) that for large n, with probability greater than $1 - \frac{1}{\bar{p}_n^3}$, there exists some positive constant $\bar{C}_2$ such that

$$\begin{aligned} &\max_{1\le i\le n} |\hat{S}(Y_i \,|\, \mathbf{X}_i) - S(Y_i \,|\, \mathbf{X}_i)| \le c_\tau \alpha \sqrt{\frac{\log \bar{p}_n}{n h_n^{d_1}}} + 4c_l h_n + c_l \bar{C}_2 \zeta_n, \\ \le& (c_\tau \alpha + 4c_l)\Big(\frac{\log \bar{p}_n}{n}\Big)^{\frac{1}{2+d_1}} + c_l \bar{C}_2 \zeta_n =: C_5\Big(\frac{\log \bar{p}_n}{n}\Big)^{\frac{1}{2+d_1}} + C_6 \zeta_n, \end{aligned} \tag{I.11}$$

where $\zeta_n = (\bar{b} + \sqrt{q_1})(q_1 \log \bar{p}_n/n)^{1/2}$. In addition, since $|\frac{\delta_i(X_{ij} - \mu_j)}{S(Y_i \,|\, \mathbf{X}_i)}| \le 2K\tau_1^{-1}$, by Hoeffding's inequality, for every $\epsilon > 0$,

$$P(|\tilde{m}_{yj} - m_{yj}| > \epsilon) \le P\big(N_y \le \frac{np_y}{2}\big) + P\big(|\tilde{m}_{yj} - m_{yj}| > \epsilon,\, N_y > \frac{np_y}{2}\big)$$

$$\le \exp\big(-\frac{3np_y}{28}\big) + \sum_{j=\lceil np_y/2\rceil}^{n} P(|\tilde{m}_{yj} - m_{yj}| > \epsilon,\, N_y = j) \le \exp\big(-\frac{3np_y}{28}\big)$$

$$+ \sum_{j=\lceil np_y/2\rceil}^{n} 2\exp\big(-\frac{j\tau_1^2\epsilon^2}{2K^2}\big) \le \exp\big(-\frac{3np_y}{28}\big) + \frac{8K^2}{\tau_1^2\epsilon^2}\exp\big(-\frac{\tau_1^2 np_y\epsilon^2}{4K^2}\big).$$

From the display above, there exists a constant $C_7 > 0$ such that by taking $\epsilon = C_7\sqrt{\frac{b\log\bar{p}_n}{n}}$, for large enough n, with probability greater than $1 - \frac{1}{\bar{p}_n^3}$,

$$|\tilde{m}_{yj} - m_{yj}| \le C_7\sqrt{\frac{b\log\bar{p}_n}{n}}. \tag{I.12}$$

Also, by Lemma I.1, with probability greater than $1 - \frac{4}{\bar{p}_n^3}$,

$$|\bar{X}_j - \mu_j| \le \sqrt{\frac{8C\log\bar{p}_n}{n}} \text{ and } |\hat{p}_y - p_y| \le \sqrt{\frac{4\log\bar{p}_n}{n}}. \tag{I.13}$$

Therefore, by (I.9)-(I.13), for large n, with probability greater than $1 - \frac{6}{\bar{p}_n^3}$

$$|\hat{m}_{2yj} - m_{2yj}| \le 8K\tau_1^{-2}\Big(C_5\big(\frac{\log\bar{p}_n}{n}\big)^{\frac{1}{2+d_1}} + C_6\zeta_n\Big) + C_7\sqrt{\frac{b\log\bar{p}_n}{n}},$$

which implies that with probability greater than $1 - \frac{6b}{\bar{p}_n^2}$,

$$|\tilde{G}_1| \le \Big(8K\tau_1^{-2}b^{1/2}\big(C_5\big(\log\bar{p}_n/n\big)^{\frac{1}{2+d_1}} + C_6\zeta_n\big) + C_7 b(\log\bar{p}_n/n)^{1/2}\Big)\sum_{j=1}^{p}\|\hat{\boldsymbol{\delta}}_{0j}\|_2. \tag{I.14}$$

Also, following arguments similar for that of (I.36), with probability greater than $1 - 1/\bar{p}_n^5$,

$$|\tilde{G}_2| \le C_2\big(\frac{q\log\bar{p}_n}{n}\big)^{1/2}\sum_{j=1}^{p}\|\hat{\boldsymbol{\delta}}_{0j}\|_2. \tag{I.15}$$

We now assume that (I.14) and (I.15) hold. Define

$$c_0 = 4K\tau_1^{-2}(C_5 + C_6) + \frac{1}{2}(C_2 + C_7). \tag{I.16}$$

By (I.8), (I.14) and (I.15), we have for large n that

$$\tilde{G}_3 + \sum_{j\in\mathcal{A}_0^c}(\lambda w_j - c_0\xi_n)\|\hat{\boldsymbol{\delta}}_{0j}\|_2 \le \sum_{j\in\mathcal{A}_0}(\lambda w_j + c_0\xi_n)\|\hat{\boldsymbol{\delta}}_{0j}\|_2,$$

which implies that, if for every $j \in \mathcal{A}_0$, $\lambda w_j \le 2c_0\xi_n$ and for every $j \in \mathcal{A}_0^c$,

$\lambda w_j \geq 2c_0\xi_n$, then

$$\sum_{j\in\mathcal{A}_0^c} c_0\xi_n\|\hat{\boldsymbol{\delta}}_{0j}\|_2 \leq \tilde{G}_3 + \frac{1}{2}\sum_{j\in\mathcal{A}_0^c} \lambda w_j\|\hat{\boldsymbol{\delta}}_{0j}\|_2 \leq 3\sum_{j\in\mathcal{A}_0} c_0\xi_n\|\hat{\boldsymbol{\delta}}_{0j}\|_2, \tag{I.17}$$

and $\sum_{j\in\mathcal{A}_0^c}\|\hat{\boldsymbol{\delta}}_{0j}\|_2 \leq 3\sum_{j\in\mathcal{A}_0}\|\hat{\boldsymbol{\delta}}_{0j}\|_2$. Define $\tilde{G}_4 = \frac{1}{2}\mathrm{tr}(\hat{\Delta}_0^T\Sigma\hat{\Delta}_0)$, and let $\tilde{\mathcal{A}}_0 = \mathcal{A}_0 \cup \hat{\mathcal{A}}_{01}$, where $\hat{\mathcal{A}}_{01}$ is the index set in $\mathcal{A}_0^c$ that corresponds to the q largest $\|\hat{\boldsymbol{\delta}}_{0j}\|_2$ with $j\in\mathcal{A}_0^c$. By (I.17) and an argument similar to (I.38),

$$\tilde{G}_4 \leq 3c_0\xi_n q^{1/2}\big(\sum_{j\in\tilde{\mathcal{A}}_0}\|\hat{\boldsymbol{\delta}}_{0j}\|_2^2\big)^{1/2} + 8C_2 q\sqrt{\frac{\log\bar{p}_n}{n}}\big(\sum_{j\in\tilde{\mathcal{A}}_0}\|\hat{\boldsymbol{\delta}}_{0j}\|_2^2\big),$$

and for large n,

$$\big(\sum_{j\in\tilde{\mathcal{A}}_0}\|\hat{\boldsymbol{\delta}}_{0j}\|_2^2\big)^{1/2} \leq \frac{3c_0q^{1/2}\xi_n}{\tilde{G}_4/(\sum_{j\in\tilde{\mathcal{A}}_0}\|\hat{\boldsymbol{\delta}}_{0j}\|_2^2) - 8C_2q\sqrt{\log\bar{p}_n/n}} \leq 12c_0\sigma_*^{-1}q^{1/2}\xi_n. \tag{I.18}$$

Similar to arguments in (I.40), we can obtain that $\sum_{j\in\mathcal{A}_0^c\setminus\hat{\mathcal{A}}_{01}}\|\hat{\boldsymbol{\delta}}_{0j}\|_2^2 \leq 9\sum_{j\in\mathcal{A}_0}\|\hat{\boldsymbol{\delta}}_{0j}\|_2^2$. In together with the display above and probability bounds for (I.14) and (I.15), we conclude that with probability greater than $1-1/\bar{p}_n$,

$$\|\hat{\Gamma}_0\hat{\Phi}_0 - \Gamma_0\Phi_0\|_F \leq 38c_0\sigma_*^{-1}q^{1/2}\xi_n. \tag{I.19}$$

Then by Wedin's theorem and (C5), we complete the proof. $\square$

Proof of Theorem 4.2. To show (4.3), with Theorem I.1, we need to verify that $\mathcal{W}_0$ occurs with high probability, where $\mathcal{W}_0 = \{\forall j\in\mathcal{A}_0, \lambda w_j \leq 2c_0\xi_n \text{ and } \forall j\in\mathcal{A}_0^c, \lambda w_j \geq 2c_0\xi_n\}$. Defining $\nu_{2n} = \min_{j\in\mathcal{A}_0}\sqrt{\mathbf{e}_j^T\tilde{U}_c\tilde{U}_c^T\mathbf{e}_j}$, for large n, by (I.17), (I.18) and Condition (C5), with probability greater than $1-1/\bar{p}_n$, for all $j\in\mathcal{A}_0$,

$$\|\tilde{\gamma}_{0j}\|_2 \geq \nu_{2n} - 12c_0\sigma_*^{-1}q^{1/2}\xi_n \geq \frac{\nu_{2n}}{2}, \tag{I.20}$$

and for all $j\in\mathcal{A}_0^c$, $\|\tilde{\gamma}_{0j}\|_2 \leq 36c_0\sigma_*^{-1}q^{1/2}\xi_n$. These results imply that for all $j\in\mathcal{A}_0$, $\lambda w_j \leq \frac{\lambda}{(\nu_{2n}/2)^\rho} \leq \frac{\lambda}{(c_{u2}^{1/2}n^{-\tau_2/2}/2)^\rho} = 2c_0\xi_n$, and for all $j\in\mathcal{A}_0^c$,

$$\lambda w_j \geq \frac{\lambda}{(36c_0\sigma_*^{-1}q^{1/2}\xi_n)^\rho} = \frac{2^{1-\rho}c_0c_{u2}^{\rho/2}n^{-\rho\tau_2/2}\xi_n}{(36c_0\sigma_*^{-1}q^{1/2}\xi_n)^\rho} \geq C_8n^{\rho(\zeta_2-\tau_2)/2}\xi_n \geq 2c_0\xi_n, \tag{I.21}$$

where C_8 is some positive constant. In together with Theorem I.1, the two displays above imply that (4.3) holds.

To show (4.4), using the same arguments as (I.20), we have that with probability greater than $1-2/\bar{p}_n$, for all $j\in\mathcal{A}_0$, $\|\hat{\gamma}_{0j}\|_2 > \nu_{2n}/2 > 0$. Therefore, $P(\mathcal{A}_0 \subseteq \hat{\mathcal{A}}_0) \to 1$ as $n\to\infty$. In addition, given $j\in\mathcal{A}_0^c$, if

$\|\hat{\boldsymbol{\gamma}}_{0j}\|_2 > 0$, by KKT conditions,

$$\|\mathbf{e}_j^T(\hat{U}_c - \Sigma_n \hat{\Gamma}_0 \hat{\Phi}_0)\hat{\Phi}_0^T\|_2 = \lambda w_j. \tag{I.22}$$

Noting that

$$\begin{aligned}\hat{U}_c - \Sigma_n \hat{\Gamma}_0 \hat{\Phi}_0 &= [\hat{U}_c - \Sigma \Gamma_0 \Phi_0] - [(\Sigma_n - \Sigma)(\hat{\Gamma}_0 \hat{\Phi}_0 - \Gamma_0 \Phi_0)] \\ &\quad - [\Sigma(\hat{\Gamma}_0 \hat{\Phi}_0 - \Gamma_0 \Phi_0)] + [(\Sigma_n - \Sigma)\Gamma_0 \Phi_0] =: A_{21} - A_{22} - A_{23} - A_{24},\end{aligned}$$

we can use arguments similar to (I.14), (I.15) and the conclusion in Theorem I.1 to obtain that, with probability greater than $1 - 6/\bar{p}_n^3$,

$$\begin{aligned}&\|\mathbf{e}_j^T A_{21}\|_2 \le 8K\tau_1^{-2} b^{1/2}\Big(C_5\big(\log \bar{p}_n/n\big)^{\frac{1}{2+d_1}} + C_6 \zeta_n\Big) + C_7 b(\log \bar{p}_n/n)^{1/2}, \\ &\|\mathbf{e}_j^T A_{22}\|_2 \le 38 c_0 C_2 \sigma_*^{-1} \xi_n (q \log \bar{p}_n/n)^{1/2}, \\ &\|\mathbf{e}_j^T A_{23}\|_2 \le 38 c_0 \tilde{\sigma} \sigma_*^{-1} q^{1/2} \xi_n, \quad \|\mathbf{e}_j^T A_{24}\|_2 \le C_2 \sqrt{q \log \bar{p}_n/n}.\end{aligned}$$

By (I.21), the displays above imply that there is a constant $C_9 > 0$ such that $\|\mathbf{e}_j^T(\hat{U}_c - \Sigma_n \hat{\Gamma}_0 \hat{\Phi}_0^T)\|_2 \le C_9 q^{1/2} \xi_n \le \lambda w_j$, which contradicts (I.22). Therefore, $P(\hat{\mathcal{A}}_0 \subseteq \mathcal{A}_0) \to 1$ as $n \to \infty$, and the proof is complete. $\square$

Proof of Theorem 4.4. Given $\Gamma_* \in \mathbb{R}^{p \times d}$, denote $\hat{Q}_*(t \,|\, \mathbf{x})$ by setting $\hat{\Gamma}_0 = \Gamma_*$ in (3.4). Let Δ_n and ϵ_n be decreasing sequences with $\Delta_n \to 0$ and $\epsilon_n \to 0$ as $n \to \infty$. Define $\mathbf{Z} = \Gamma_*^T \mathbf{X}$, $\tilde{\mathbf{Z}} = \Gamma_0^T \mathbf{X}$, $\mathbf{z}_0 = \Gamma_*^T \mathbf{x}_0$ and $\mathbf{Z}_i = \Gamma_*^T \mathbf{X}_i$. Define $\hat{f}_1(t \,|\, \mathbf{z}_0) = \hat{Q}_*(t \,|\, \mathbf{x}_0)$ and $f_1(t \,|\, \mathbf{z}_0) = \frac{1}{M_h}\int_{\mathbb{R}^d} P(T > t \,|\, \mathbf{Z} = \mathbf{z}) K(\frac{\mathbf{z} - \mathbf{z}_0}{h_n}) p_{\mathbf{Z}}(\mathbf{z}) d\mathbf{z}$, where $M_h = \int_{\mathbb{R}^d} K(\frac{\mathbf{z} - \mathbf{z}_0}{h_n}) p_{\mathbf{Z}}(\mathbf{z}) d\mathbf{z}$ and $p_{\mathbf{Z}}(\mathbf{z})$ is the density of $\mathbf{Z}$. Similarly, define $f_2(t \,|\, \mathbf{z}_0) = \frac{1}{\tilde{M}_h}\int_{\mathbb{R}^d} P(T > t \,|\, \tilde{\mathbf{Z}} = \mathbf{z}) K(\frac{\mathbf{z} - \mathbf{z}_0}{h_n}) p_{\tilde{\mathbf{Z}}}(\mathbf{z}) d\mathbf{z}$, where $\tilde{M}_h = \int_{\mathbb{R}^d} K(\frac{\mathbf{z} - \mathbf{z}_0}{h_n}) p_{\tilde{\mathbf{Z}}}(\mathbf{z}) d\mathbf{z}$ and $p_{\tilde{\mathbf{Z}}}(\mathbf{z})$ is the density of $\tilde{\mathbf{Z}}$. Assume for now that $\|\Gamma_*^T \mathbf{x} - \Gamma_0^T \mathbf{x}\|_F \le \Delta_n$ for every $\mathbf{x} \in \mathcal{X}$. Then by Condition (C2), the Lipschitz condition holds for $m_0(t \,|\, \mathbf{z}) := P(T > t \,|\, \Gamma_0^T \mathbf{X} = \mathbf{z})$, which implies

$$\begin{aligned}|\hat{Q}_*(t \,|\, \mathbf{x}_0) - Q(t \,|\, \mathbf{x}_0)| &\le \big|\hat{f}_1(t \,|\, \mathbf{z}_0) - f_2(t \,|\, \mathbf{z}_0)\big| + \big|f_2(t \,|\, \mathbf{z}_0) - m_0(t \,|\, \mathbf{z}_0)\big| \\ &\le \big|\hat{f}_1(t \,|\, \mathbf{z}_0) - f_2(t \,|\, \mathbf{z}_0)\big| + c_l h_n.\end{aligned} \tag{I.23}$$

In addition, note that

$$\big|\hat{f}_1(t \,|\, \mathbf{z}_0) - f_2(t \,|\, \mathbf{z}_0)\big| \le \big|\hat{f}_1(t \,|\, \mathbf{z}_0) - f_1(t \,|\, \mathbf{z}_0)\big| + \big|f_1(t \,|\, \mathbf{z}_0) - f_2(t \,|\, \mathbf{z}_0)\big|, \tag{I.24}$$

where

$$\begin{aligned}\big|f_1(t \,|\, \mathbf{z}_0) - f_2(t \,|\, \mathbf{z}_0)\big| &\le \big|f_1(t \,|\, \mathbf{z}_0) - \tilde{f}(t \,|\, \mathbf{z}_0)\big| + \big|\tilde{f}(t \,|\, \mathbf{z}_0) - f_2(t \,|\, \mathbf{z}_0)\big| \\ &\le c_l \Delta_n + c_l h_n,\end{aligned} \tag{I.25}$$

with $\tilde{f}(t \,|\, \mathbf{z}_0) = \frac{1}{M_h}\int_{\mathbb{R}^d} P(T > t \,|\, \tilde{\mathbf{Z}} = \mathbf{z}) K(\frac{\mathbf{z} - \mathbf{z}_0}{h_n}) p_{\mathbf{Z}}(\mathbf{z}) d\mathbf{z}$, and the last inequality in (I.25) follows by noting that $f_1(t \,|\, \mathbf{z}_0)$ and $\tilde{f}(t \,|\, \mathbf{z}_0)$ can be represented by

$$f_1(t \mid \mathbf{z}_0) = \frac{1}{M_h}\int_{\mathcal{X}} P(T > t \mid \mathbf{X} = \mathbf{x}) p_{\mathbf{X}}(\mathbf{x}) K\big(\frac{\Gamma_*^T\mathbf{x} - \Gamma_*^T\mathbf{x}_0}{h_n}\big) d\mathbf{x},$$
$$\tilde{f}(t \mid \mathbf{z}_0) = \frac{1}{M_h}\int_{\mathcal{X}} P(T > t \mid \Gamma_0^T\mathbf{X} = \Gamma_*^T\mathbf{x}) p_{\mathbf{X}}(\mathbf{x}) K\big(\frac{\Gamma_*^T\mathbf{x} - \Gamma_*^T\mathbf{x}_0}{h_n}\big) d\mathbf{x}.$$

Then by (I.23), (I.24) and (I.25),

(I.26)
$$\sup_{t\in[0,T_0)} |\hat{Q}_*(t \mid \mathbf{x}_0) - Q(t \mid \mathbf{x}_0)| \leq \sup_{t\in[0,T_0)} \big|\hat{f}_1(t \mid \mathbf{z}_0) - f_1(t \mid \mathbf{z}_0)\big| + c_l\Delta_n + 2c_l h_n.$$

We next focus on studying $\sup_{t\in[0,T_0)}\big|\hat{f}_1(t \mid \mathbf{z}_0) - f_1(t \mid \mathbf{z}_0)\big|$. Define $H_1(t \mid \mathbf{z}_0) = \frac{1}{M_h}\int_{\mathbb{R}^d} P(Y > t, \delta = 1 \mid \mathbf{Z} = \mathbf{z}) K(\frac{\mathbf{z}-\mathbf{z}_0}{h_n}) p_{\mathbf{Z}}(\mathbf{z}) d\mathbf{z}$ and define $H_2(t \mid \mathbf{z}_0) = \frac{1}{M_h}\int_{\mathbb{R}^d} P(Y > t \mid \mathbf{Z} = \mathbf{z}) K(\frac{\mathbf{z}-\mathbf{z}_0}{h_n}) p_{\mathbf{Z}}(\mathbf{z}) d\mathbf{z}$. Inspired by Dabrowska (1989), define $\Lambda(t \mid \mathbf{z}_0) := -\int_0^t \frac{dH_1(s \mid \mathbf{z}_0)}{H_2(s \mid \mathbf{z}_0)}$ similar to cumulative hazard function. Note that by Condition (C1), for every $t \in [0, T_0)$, $f_1(t \mid \mathbf{z}_0) \geq \tau_0$ and $H_2(t \mid \mathbf{z}_0) \geq \tau_0^2 =: \tau_1$. Correspondingly, define $\hat{H}_1(t \mid \mathbf{z}_0) = \frac{\sum_{i=1}^n I(Y_i > t, \delta_i = 1) K\left(\frac{\mathbf{Z}_i - \mathbf{z}_0}{h_n}\right)}{\sum_{i=1}^n K\left(\frac{\mathbf{Z}_i - \mathbf{z}_0}{h_n}\right)}$, $\hat{H}_2(t \mid \mathbf{z}_0) = \frac{\sum_{i=1}^n I(Y_i > t) K\left(\frac{\mathbf{Z}_i - \mathbf{z}_0}{h_n}\right)}{\sum_{i=1}^n K\left(\frac{\mathbf{Z}_i - \mathbf{z}_0}{h_n}\right)}$ and

$$\hat{\Lambda}(t \mid \mathbf{z}_0) = -\int_0^t \frac{d\hat{H}_1(s \mid \mathbf{z}_0)}{\hat{H}_{2j}(s- \mid \mathbf{z}_0)} = \sum_{k=1}^n \frac{I(Y_k \leq t, \delta_k = 1) K\left(\frac{\mathbf{Z}_k - \mathbf{z}_0}{h_n}\right)}{\sum_{i=1}^n I(Y_i > Y_k) K\left(\frac{\mathbf{Z}_i - \mathbf{z}_0}{h_n}\right)}.$$

Then by definition of $\hat{f}_1(t \mid \mathbf{z}_0)$,

$$\hat{f}_1(t \mid \mathbf{z}_0) = \prod_{k=1}^n \Big\{1 - \frac{I(Y_k \leq t, \delta_k = 1) K\left(\frac{\mathbf{Z}_k - \mathbf{z}_0}{h_n}\right)}{\sum_{i=1}^n I(Y_i \geq Y_k) K\left(\frac{\mathbf{Z}_i - \mathbf{z}_0}{h_n}\right)}\Big\} = \prod_{s \leq t}\{1 - d\hat{\Lambda}(s \mid \mathbf{z}_0)\},$$

where $\prod$ is the product integral (Gill & Johansen 1990). By arguments similar to that of (I.53), for every $t \in [0, T_0)$,

$$\begin{aligned} &|\hat{f}_1(t \mid \mathbf{z}_0) - f_1(t \mid \mathbf{z}_0)| \\ \leq &\Big|\int_0^t \frac{\hat{f}_1(s- \mid \mathbf{z}_0)}{f_1(s \mid \mathbf{z}_0)} d[\hat{\Lambda}(s \mid \mathbf{z}_0) - \Lambda(s \mid \mathbf{z}_0)]\Big| \leq \frac{3}{\tau_0}|\hat{\Lambda}(t \mid \mathbf{z}_0) - \Lambda(t \mid \mathbf{z}_0)|. \end{aligned} \tag{I.27}$$

With function $f(t)$ and $T > 0$, define $\|f\|_T = \sup_{t\in[0,T)} f(t)$. Note that under the event $\mathcal{G}_0 := \{\|\hat{H}_2(\cdot \mid \mathbf{z}_0) - H_2(\cdot \mid \mathbf{z}_0)\|_{T_0} \leq \tau_1/2\}$, with arguments similar to that of (I.54),

$$\begin{aligned} &|\hat{\Lambda}(t \mid \mathbf{z}_0) - \Lambda(t \mid \mathbf{z}_0)| \\ \leq &8\tau_1^{-2}\|\hat{H}_2(\cdot \mid \mathbf{z}_0) - H_2(\cdot \mid \mathbf{z}_0)\|_{T_0} + 2\tau_1^{-1}\|\hat{H}_1(\cdot \mid \mathbf{z}_0) - H_1(\cdot \mid \mathbf{z}_0)\|_{T_0}. \end{aligned}$$

Then under $\mathcal{G}_0$, the two displays above imply that

$$B_0 := \sup_{t\in[0,T_0)} |\hat{f}_1(t\,|\,\mathbf{z}_0) - f_1(t\,|\,\mathbf{z}_0)|$$

$$\leq \frac{24}{\tau_0\tau_1^2}\|\hat{H}_2(\cdot\,|\,\mathbf{z}_0) - H_2(\cdot\,|\,\mathbf{z}_0)\|_{T_0} + \frac{6}{\tau_0\tau_1}\|\hat{H}_1(\cdot\,|\,\mathbf{z}_0) - H_1(\cdot\,|\,\mathbf{z}_0)\|_{T_0}. \tag{I.28}$$

Next, we want to provide upper bound for B_0. Given domain $\mathcal{Z}$ of $\mathbf{Z}$ and $r > 0$, define the set of interior points $\mathcal{Z}_r = \{\mathbf{z} \in \mathcal{Z} :\, B_2(\mathbf{z}, r) \subset \mathcal{Z}\}$, where $B_2(\mathbf{z}, r)$ is the l_2-ball centered at $\mathbf{z}$ with radius r. Given h_n with $h_n \to 0$, define $\eta = h_n/4$. Let $\mathcal{Q}_\eta$ be the η-packing set of $\mathcal{Z}_\eta$ in l_2-norm with highest possible cardinality (i.e., $|\mathcal{Q}_\eta|$ gives the maximal number of $\frac{\eta}{2}$-balls in l_2-norm that we possibly pack in $\mathcal{Z}_\eta$; Yang & Barron 1999). Let $\mathbf{z}_*$ be the closest point to $\mathbf{z}_0$ in $\mathcal{Z}_\eta$ so that $B_2(\mathbf{z}_*, \eta) \subset B_2(\mathbf{z}_0, h_n)$. Define N_* and N_0 be the number of sample points in $B_2(\mathbf{z}_*, \eta)$ and $B_2(\mathbf{z}_0, h_n)$, respectively. Then, by the extended Bernstein inequality (Lemma 2, Qian & Yang 2016), there is a constant $\tilde{c}_u > 0$ such that

$$P\big(N_0 \leq \frac{n\tilde{c}_u h_n^d}{2}\big) \leq P\big(N_* \leq \frac{n\tilde{c}_u h_n^d}{2}\big) \leq \exp\big(-\frac{3\tilde{c}_u n h_n^d}{28}\big). \tag{I.29}$$

Suppose $\epsilon_n \to 0$ and $n^\nu \epsilon_n \to \infty$ for some $\nu \geq 1/2$. Then, note by (I.5) in Lemma I.3 that there is a constant $\bar{\alpha} > 0$ such that for every $j \geq \frac{n\tilde{c}_u h_n^d}{2}$, if $\epsilon_n \geq \bar{\alpha}\sqrt{\frac{\log n}{nh_n^d}}$, then $\epsilon_n \geq 2\sqrt{\frac{\log N_{\epsilon_n}}{j}}$, where N_{ϵ_n} follows the definition in Lemma I.4. Indeed, for any $\bar{\alpha} \geq 4\sqrt{\nu a_1/\tilde{c}_u}$ and large enough n with a_1 from Lemma I.3, we have

$$j\epsilon_n^2 \geq \frac{n\tilde{c}_u h_n^d \epsilon_n^2}{2} \geq \frac{\bar{\alpha}^2\tilde{c}_u \log n}{2} \geq \frac{\bar{\alpha}^2\tilde{c}_u}{2\nu}\log\big(\frac{1}{\epsilon_n}\big) \geq \frac{\bar{\alpha}^2\tilde{c}_u}{4\nu a_1}\log N_{\epsilon_n} \geq 4\log N_{\epsilon_n}. \tag{I.30}$$

Then, defining $\tilde{n}_1 = \lceil \frac{n\tilde{c}_u h_n^d}{2} \rceil$, with our choice of ϵ_n, by (I.29) and Lemma I.4, we have

$$\begin{aligned} r_{20} :=& P(\|\hat{H}_2(\cdot\,|\,\mathbf{z}_0) - H_2(\cdot\,|\,\mathbf{z}_0)\|_{T_0} > 8\epsilon_n) \\ \leq& P\big(N_0 \leq \frac{n\tilde{c}_u h_n^d}{2}\big) + \sum_{k=\tilde{n}_1}^{n} P(\|\hat{H}_2(\cdot\,|\,\mathbf{z}_0) - H_2(\cdot\,|\,\mathbf{z}_0)\|_{T_0} > 8\epsilon_n,\, N_0 = k) \\ \leq& \exp\big(-\frac{3\tilde{c}_u n h_n^d}{28}\big) + \sum_{k=\tilde{n}_1}^{n} 8\exp\big(-\frac{j\epsilon_n^2}{4}\big) \\ \leq& \exp\big(-\frac{3\tilde{c}_u n h_n^d}{28}\big) + \frac{64}{\epsilon_n^2}\exp\big(-\frac{\tilde{n}_1\epsilon_n^2}{4}\big) =: \kappa_n. \end{aligned} \tag{I.31}$$

Similarly, we can show that

$$r_{10} := P(\|\hat{H}_1(\cdot \,|\, \mathbf{z}_0) - H_1(\cdot \,|\, \mathbf{z}_0)\|_{T_0} > 8\epsilon_n) \le \kappa_n. \tag{I.32}$$

Consequently, by (I.26), (I.28), (I.31), (I.32) and noting that

$$P(B_0 \ge c_\tau \epsilon_n) \le P(\mathcal{G}_0^c) + P(B_0 \ge c_\tau \epsilon_n,\, \mathcal{G}_0) \le P(\mathcal{G}_0^c) + r_{20} + r_{10} \le 3\kappa_n,$$

with probability great than $1 - \bar{c}_{11}\exp(-\bar{c}_{12} n h_n^d) - \bar{c}_{21}\epsilon_n^{-2}\exp(-\bar{c}_{22} n h_n^d \epsilon_n^2)$, we have $\sup_{t\in[0,T_0)} |\hat{Q}_*(t \,|\, \mathbf{x}_0) - Q(t \,|\, \mathbf{x}_0)| \le \bar{c}_\tau \epsilon_n + 4c_l h_n + c_l \Delta_n$, where $\bar{c}_\tau = 192\tau_0^{-1}\tau_1^{-2} + 48\tau_0^{-1}\tau_1^{-1}$, $\bar{c}_{11} = 3$, $\bar{c}_{12} = 3\tilde{c}_u/28$, $\bar{c}_{21} = 192$ and $\bar{c}_{22} = \tilde{c}_u/8$. In together with (I.19), Wedin's theorem and setting $\epsilon_n = \bar{\alpha}\sqrt{\frac{\log n}{n h_n^d}}$ with a (large) constant $\bar{\alpha} > 0$, we obtain (I.60). This completes the proof of Theorem 4.4. Corollary 4.1 then follows immediately from Lemma 21.1 of van der Vaart (2000) and the uniform consistency of $\hat{Q}(t \,|\, \mathbf{x}_0)$. □

I.4. *Proof of Theorems 4.1 and 4.3.*

THEOREM I.2. *Suppose Conditions (A1)-(A7) hold, and assume that for every $j \in \mathcal{A}_1$, $\lambda_1 w_j \le 2c_{1n}\sqrt{\frac{\log \bar{p}_n}{n}}$ and for every $j \in \mathcal{A}_1^c$, $\lambda_1 w_j \ge 2c_{1n}\sqrt{\frac{\log \bar{p}_n}{n}}$, where*

$$c_{1n} = (C_1 \bar{b} + C_2 q_1^{1/2})/2, \tag{I.33}$$

and C_1 and C_2 are some generic positive constants given in (I.35) and (I.36). Then there exist constants $C_3, C_4 > 0$ such that with probability great than $1 - C_3/\bar{p}_n^4$,

$$\|\mathrm{P}_{\hat{\Gamma}_1} - \mathrm{P}_{\Gamma_1}\|_F \le C_4(\bar{b} + q_1^{1/2})\sqrt{\frac{q_1 \log \bar{p}_n}{n}}.$$

PROOF OF THEOREM I.2. Recall that Γ_1 is a basis of $\mathcal{S}_{(T,C)|\mathbf{X}}$ and $(\hat{\Gamma}_1, \hat{\Phi}_1)$ is the minimizer of (3.6). suppose $\tilde{U}_1 = \Gamma_1 \Phi_1$, where Φ_1 is the coordinate matrix. Since $F_{1n}(\hat{\Gamma}_1, \hat{\Phi}_1) \le F_{1n}(\Gamma_1, \Phi_1)$,

$$\begin{aligned} &- \mathrm{tr}(\hat{U}_1^T \hat{\Gamma}_1 \hat{\Phi}_1) + \frac{1}{2}\mathrm{tr}(\hat{\Phi}_1^T \hat{\Gamma}_1^T \Sigma_n \hat{\Gamma}_1 \hat{\Phi}_1) + \lambda_1 \sum_{j=1}^p w_j \|\hat{\boldsymbol{\gamma}}_{1j}\|_2 \\ &\le -\mathrm{tr}(\hat{U}_1^T \Gamma_1 \Phi_1) + \frac{1}{2}\mathrm{tr}(\Phi_1^T \Gamma_1^T \Sigma_n \Gamma_1 \Phi_1) + \lambda_1 \sum_{j=1}^p w_j \|\boldsymbol{\gamma}_{1j}\|_2, \end{aligned}$$

which implies that

$$
\begin{aligned}
&G_1 + G_2 + G_3 + \lambda_1 \sum_{j=1}^{p} w_j \|\hat{\boldsymbol{\gamma}}_{1j}\|_2 \\
:=& -\operatorname{tr}[(\hat{U}_1 - \Sigma\Gamma_1\Phi_1)^T \hat{\Delta}_1] + \operatorname{tr}[\Phi_1^T \Gamma_1^T (\Sigma_n - \Sigma)\hat{\Delta}_1] \\
&+ \frac{1}{2}\operatorname{tr}(\hat{\Delta}_1^T \Sigma_n \hat{\Delta}_1) + \lambda_1 \sum_{j=1}^{p} w_j \|\hat{\boldsymbol{\gamma}}_{1j}\|_2 \\
\le& \lambda_1 \sum_{j=1}^{p} w_j \|\boldsymbol{\gamma}_{1j}\|_2, \qquad \text{(I.34)}
\end{aligned}
$$

where $\hat{\Delta}_1 := \hat{\Gamma}_1\hat{\Phi}_1 - \Gamma_1\Phi_1 = (\hat{\boldsymbol{\delta}}_{11}, \cdots, \hat{\boldsymbol{\delta}}_{1p})^T$. Given $y = (k, l)$ with $k = 0, 1$ and $l = 1, \cdots, b_k$, define $\bar{X}_{yj}$ to be the sample estimator of $\mu_{yj} := \mathrm{E}(X_j \mid Y \in \mathcal{H}_y,\ \delta = k)$. Noting that with slice $\mathcal{H}_y$ and $1 \le j \le p$, we have

$$
\begin{aligned}
&|(\bar{X}_{yj} - \bar{X}_j)\sqrt{\hat{p}_y} - \mu_{yj}\sqrt{p_y}| \\
\le& |(\bar{X}_{yj} - \mu_{yj})\sqrt{\hat{p}_y}| + |\mu_{yj}(\sqrt{\hat{p}_y} - \sqrt{p_y})| + |(\bar{X}_j - \mu_j)\sqrt{\hat{p}_y}|,
\end{aligned}
$$

which implies by Lemmas I.1, I.2 and Condition (A2) that there exists a constant $C_1 > 0$ such that for large enough n, with probability greater than $1 - 1/\bar{p}_n^6$, $|(\bar{X}_{yj} - \bar{X}_j)\sqrt{\hat{p}_y} - \mu_{yj}\sqrt{p_y}| \le C_1\sqrt{\bar{b}\log \bar{p}_n / n}$. Therefore, by union bound, with probability greater than $1 - \bar{b}/\bar{p}_n^5$,

$$
|G_1| \le C_1 \bar{b} \big(\frac{\log \bar{p}_n}{n}\big)^{1/2} \sum_{j=1}^{p} \|\hat{\boldsymbol{\delta}}_{1j}\|_2. \qquad \text{(I.35)}
$$

Next, define $\hat{\sigma}_m = \max_{1 \le i \le j \le p} |\hat{\sigma}_{ij} - \sigma_{ij}|$, where $\hat{\sigma}_{ij}$ and σ_{ij} are corresponding (i, j)-element in Σ_n and Σ, respectively. Thus, by Condition (A3), we have a known result (see, e.g., Bickel & Levina 2008): there is a constant $\tilde{C}_4 > 0$ so that with probability greater than $1 - 1/\bar{p}_n^5$, $\hat{\sigma}_m \le \tilde{C}_4 \sqrt{\frac{\log \bar{p}_n}{n}}$. In together with von Neumann's trace inequality, we have

$$
\|\mathbf{e}_j^T (\Sigma_n - \Sigma)\Gamma_1\Phi_1\|_2 \le \hat{\sigma}_m \sum_{j \in \mathcal{A}_1} \|\boldsymbol{\gamma}_{1j}^T \Phi_1\|_2 \le \hat{\sigma}_m \sqrt{q_1} \|\Gamma_1\Phi_1\|_F \le \frac{\hat{\sigma}_m \sqrt{d_1 q_1}}{\sigma_*^{1/2}},
$$

which implies that

(I.36)

$$
|G_2| \le \tilde{C}_4 \sigma_*^{-1/2} \big(\frac{d_1 q_1 \log \bar{p}_n}{n}\big)^{1/2} \sum_{j=1}^{p} \|\hat{\boldsymbol{\delta}}_{1j}\|_2 \le C_2 \big(\frac{q_1 \log \bar{p}_n}{n}\big)^{1/2} \sum_{j=1}^{p} \|\hat{\boldsymbol{\delta}}_{1j}\|_2,
$$

where $C_2 = \tilde{C}_4 c_d^{1/2} \sigma_*^{-1/2}$ and c_d is an upper bound of d_1.

Next, we assume both (I.35) and (I.36) hold. Then by (I.34),

$$
\begin{aligned}
&G_3 + \lambda_1 \sum_{j\in\mathcal{A}_1^c} w_j \|\hat{\boldsymbol{\delta}}_{1j}\|_2 \\
\le &\Big(C_1\bar{b}\sqrt{\frac{\log \bar{p}_n}{n}} + C_2\sqrt{\frac{q_1\log \bar{p}_n}{n}}\Big)\sum_{j=1}^{p}\|\hat{\boldsymbol{\delta}}_{1j}\|_2 + \lambda_1 \sum_{j\in\mathcal{A}_1} w_j \|\hat{\boldsymbol{\delta}}_{1j}\|_2.
\end{aligned}
$$

With (I.33), define $\mathcal{W}_1$ to be the event that for every $j\in\mathcal{A}_1$, $\lambda_1 w_j \le 2c_{1n}\sqrt{\log \bar{p}_n/n}$, and for every $j\in\mathcal{A}_1^c$, $\lambda_1 w_j \ge 2c_{1n}\sqrt{\log \bar{p}_n/n}$. Then we have from above that

$$
G_3 + \sum_{j\in\mathcal{A}_1^c}\big(\lambda_1 w_j - c_{1n}\sqrt{\frac{\log \bar{p}_n}{n}}\big)\|\hat{\boldsymbol{\delta}}_{1j}\|_2 \le \sum_{j\in\mathcal{A}_1}\big(\lambda_1 w_j + c_{1n}\sqrt{\frac{\log \bar{p}_n}{n}}\big)\|\hat{\boldsymbol{\delta}}_{1j}\|_2,
$$

which implies under $\mathcal{W}_1$ that

(I.37)

$$
\sum_{j\in\mathcal{A}_1^c} c_{1n}\sqrt{\frac{\log \bar{p}_n}{n}}\|\hat{\boldsymbol{\delta}}_{1j}\|_2 \le G_3 + \frac{1}{2}\sum_{j\in\mathcal{A}_1^c}\lambda_1 w_j\|\hat{\boldsymbol{\delta}}_{1j}\|_2 \le 3\sum_{j\in\mathcal{A}_1} c_{1n}\sqrt{\frac{\log \bar{p}_n}{n}}\|\hat{\boldsymbol{\delta}}_{1j}\|_2,
$$

and

$$
\sum_{j\in\mathcal{A}_1^c}\|\hat{\boldsymbol{\delta}}_{1j}\|_2 \le 3\sum_{j\in\mathcal{A}_1}\|\hat{\boldsymbol{\delta}}_{1j}\|_2.
$$

The display above implies that

$$
\begin{aligned}
&|\hat{\Delta}_1^T(\Sigma - \Sigma_n)\hat{\Delta}_1| \\
\le &\hat{\sigma}_m \sum_{j\in\mathcal{A}_1}\sum_{j\in\mathcal{A}_1}|\hat{\boldsymbol{\delta}}_{1j}^T\hat{\boldsymbol{\delta}}_{1j}| + \hat{\sigma}_m \sum_{j\in\mathcal{A}_1^c}\sum_{j\in\mathcal{A}_1^c}|\hat{\boldsymbol{\delta}}_{1j}^T\hat{\boldsymbol{\delta}}_{1j}| + 2\hat{\sigma}_m \sum_{j\in\mathcal{A}_1}\sum_{j\in\mathcal{A}_1^c}|\hat{\boldsymbol{\delta}}_{1j}^T\hat{\boldsymbol{\delta}}_{1j}| \\
\le &\hat{\sigma}_m\Big(\big(\sum_{j\in\mathcal{A}_1}\|\hat{\boldsymbol{\delta}}_{1j}\|_2\big)^2 + \big(\sum_{j\in\mathcal{A}_1^c}\|\hat{\boldsymbol{\delta}}_{1j}\|_2\big)^2 + 2\big(\sum_{j\in\mathcal{A}_1}\|\hat{\boldsymbol{\delta}}_{1j}\|_2\big)\big(\sum_{j\in\mathcal{A}_1^c}\|\hat{\boldsymbol{\delta}}_{1j}\|_2\big)\Big)
\end{aligned}
$$

(I.38)

$$
\le 16C_2\sqrt{\frac{\log \bar{p}_n}{n}}\big(\sum_{j\in\mathcal{A}_1}\|\hat{\boldsymbol{\delta}}_{1j}\|_2\big)^2.
$$

We then define $\tilde{\mathcal{A}}_1 = \mathcal{A}_1 \cup \hat{\mathcal{A}}_{11}$, where $\hat{\mathcal{A}}_{11}$ is the index set in $\mathcal{A}_1^c$ that corresponds to the q_1 largest $\|\hat{\boldsymbol{\delta}}_{1j}\|_2$ with $j\in\mathcal{A}_1^c$. Define $G_4 = \frac{1}{2}\mathrm{tr}(\hat{\Delta}_1^T\Sigma\hat{\Delta}_1)$. Then (I.37) and (I.38) imply that

$$
\begin{aligned}
G_4 &\le 3c_{1n}\sqrt{\frac{\log \bar{p}_n}{n}}\sum_{j\in\mathcal{A}_1}\|\hat{\boldsymbol{\delta}}_{1j}\|_2 + 8C_2\sqrt{\frac{\log \bar{p}_n}{n}}\big(\sum_{j\in\mathcal{A}_1}\|\hat{\boldsymbol{\delta}}_{1j}\|_2\big)^2 \\
&\le 3c_{1n}\sqrt{\frac{q_1\log \bar{p}_n}{n}}\big(\sum_{j\in\tilde{\mathcal{A}}_1}\|\hat{\boldsymbol{\delta}}_{1j}\|_2^2\big)^{1/2} + 8C_2 q_1\sqrt{\frac{\log \bar{p}_n}{n}}\big(\sum_{j\in\tilde{\mathcal{A}}_1}\|\hat{\boldsymbol{\delta}}_{1j}\|_2^2\big),
\end{aligned}
$$

and consequently, for large enough n,

$$\Big(\sum_{j\in\tilde{\mathcal{A}}_1}\|\hat{\boldsymbol\delta}_{1j}\|_2^2\Big)^{1/2} \le \frac{3c_{1n}\sqrt{q_1\log\bar p_n/n}}{G_4/(\sum_{j\in\tilde{\mathcal{A}}_1}\|\hat{\boldsymbol\delta}_{1j}\|_2^2)-8C_2q_1\sqrt{\log\bar p_n/n}}$$
$$\le \frac{6c_{1n}\sqrt{q_1\log\bar p_n/n}}{\sigma_*-16C_2q_1\sqrt{\log\bar p_n/n}}$$
$$\le 12c_{1n}\sigma_*^{-1}\sqrt{q_1\log\bar p_n/n}. \tag{I.39}$$

In addition, note that

$$\sum_{j\in\mathcal{A}_1^c\setminus\hat{\mathcal{A}}_{11}}\|\hat{\boldsymbol\delta}_{1j}\|_2^2 \le \sum_{k=q_1+1}^{p-q_1}\frac{1}{k^2}\Big(\sum_{j\in\mathcal{A}_1^c}\|\hat{\boldsymbol\delta}_{1j}\|_2\Big)^2 \le \frac{1}{q_1}\Big(\sum_{j\in\mathcal{A}_1^c}\|\hat{\boldsymbol\delta}_{1j}\|_2\Big)^2 \le 9\sum_{j\in\mathcal{A}_1}\|\hat{\boldsymbol\delta}_{1j}\|_2^2. \tag{I.40}$$

The two displays above in together with the probability bounds for (I.35) and (I.36) give that

$$\|\hat\Gamma_1\hat\Phi_1-\Gamma_1\Phi_1\|_F \le 38c_{1n}\sigma_*^{-1}\sqrt{q_1\log\bar p_n/n} \tag{I.41}$$

holds with probability great than $1-C_3/\bar p_n^4$ for some positive constant C_3. Then by Wedin's theorem and (A5), we obtain the conclusion of Theorem I.2. □

Proof of Theorem 4.1. Define $\tilde\Gamma_1=(\tilde{\boldsymbol\gamma}_{11}^T,\cdots,\tilde{\boldsymbol\gamma}_{1p}^T)^T$ to be the initial DS estimator corresponding to the equal penalty weights. Take $w_j=\frac{1}{\|\tilde{\boldsymbol\gamma}_{1j}\|_2^\rho}$ and define event $\mathcal{W}_1=\{\forall j\in\mathcal{A}_1,\ \lambda_1w_j\le 2c_{1n}\sqrt{\frac{\log\bar p_n}{n}}$ and $\forall j\in\mathcal{A}_1^c,\ \lambda_1w_j\ge 2c_{1n}\sqrt{\frac{\log\bar p_n}{n}}\}$. We first want to verify that $\mathcal{W}_1$ holds with high probability. Define $\nu_{1n}=\min_{j\in\mathcal{A}_1}\sqrt{\mathbf{e}_j^T\tilde U_1\tilde U_1^T\mathbf{e}_j}$. By Condition (A6), (I.37), (I.39) and our choice of $\tilde\lambda_1$, with probability great than $1-C_3/\bar p_n^4$, for all $j\in\mathcal{A}_1$,

$$\|\tilde{\boldsymbol\gamma}_{1j}\|_2\ge\nu_{1n}-12c_{1n}\sigma_*^{-1}\sqrt{q_1\log\bar p_n/n}>\nu_{1n}/2, \tag{I.42}$$

and for all $j\in\mathcal{A}_1^c$,

$$\|\tilde{\boldsymbol\gamma}_{1j}\|_2\le 36c_{1n}\sigma_*^{-1}\sqrt{q_1\log\bar p_n/n}.$$

With above, for all $j\in\mathcal{A}_1$,

$$\lambda_1w_j\le\frac{\lambda_1}{(\nu_{1n}/2)^\rho}\le\frac{\lambda_1}{(c_{u1}^{1/2}n^{-\tau_1/2}/2)^\rho}=2c_{1n}\sqrt{\log\bar p_n/n},$$

and for all $j \in \mathcal{A}_1^c$,

$$\begin{aligned} \lambda_1 w_j &\geq \frac{\lambda_1}{(36c_{1n}\sigma_*^{-1}\sqrt{q_1 \log \bar{p}_n/n})^\rho} = \frac{2^{1-\rho} c_{u1}^{\rho/2} c_{1n}\sqrt{\log \bar{p}_n/n^{1+\rho\tau_1}}}{(36c_{1n}\sigma_*^{-1}\sqrt{q_1 \log \bar{p}_n/n})^\rho} \\ &\geq C_7 n^{\rho(\zeta_1-\tau_1)/2} c_{1n}\sqrt{\log \bar{p}_n/n} \qquad \text{(I.43)} \\ &\geq 2c_{1n}\sqrt{\log \bar{p}_n/n}, \end{aligned}$$

where C_7 is a positive constant. Therefore, (4.1) holds from Theorem I.2.

To show (4.2), using same arguments as (I.42), we obtain that with probability greater than $1-2C_3/\bar{p}_n^4$, for all $j \in \mathcal{A}_1$, $\|\hat{\gamma}_{1j}\|_2 > \nu_{1n}/2 > 0$. Therefore, $P(\mathcal{A}_1 \subseteq \hat{\mathcal{A}}_1) \to 1$ as $n \to \infty$. In addition, given $j \in \mathcal{A}_1^c$, if $\|\hat{\gamma}_{1j}\|_2 > 0$, by KKT conditions,

$$\|\mathbf{e}_j^T(\hat{U}_1 - \Sigma_n\hat{\Gamma}_1\hat{\Phi}_1)\hat{\Phi}_1^T\|_2 = \lambda_1 w_j. \qquad \text{(I.44)}$$

Note that

$$\begin{aligned} &\hat{U}_1 - \Sigma_n\hat{\Gamma}_1\hat{\Phi}_1 \\ =&[\hat{U}_1 - \Sigma\Gamma_1\Phi_1] - [(\Sigma_n - \Sigma)(\hat{\Gamma}_1\hat{\Phi}_1 - \Gamma_1\Phi_1)] \\ &- [\Sigma(\hat{\Gamma}_1\hat{\Phi}_1 - \Gamma_1\Phi_1)] + [(\Sigma_n - \Sigma)\Gamma_1\Phi_1] \\ =:&A_{11} - A_{12} - A_{13} - A_{14}. \end{aligned}$$

Then, using similar arguments as that of (I.35), (I.36) and the conclusion in Theorem I.2, with probability greater than $1 - C_3/\bar{p}_n^4$, for all $j \in \mathcal{A}_1^c$,

$$\begin{aligned} &\|\mathbf{e}_j^T A_{11}\|_2 \leq C_1\bar{b}\sqrt{\frac{\log \bar{p}_n}{n}}, \\ &\|\mathbf{e}_j^T A_{12}\|_2 \leq \hat{\sigma}_m \sum_{j=1}^p \|\hat{\boldsymbol{\delta}}_{1j}\|_2 \leq 38C_2 c_{1n}\sigma_*^{-1} q_1 \log \bar{p}_n/n, \\ &\|\mathbf{e}_j^T A_{13}\|_2 \leq \tilde{\sigma}\|\hat{\Gamma}_1\hat{\Phi}_1 - \Gamma_1\Phi_1\|_F \leq 38\sigma_*^{-1}\tilde{\sigma} c_{1n}\sqrt{q_1 \log \bar{p}_n/n}, \\ &\|\mathbf{e}_j^T A_{14}\|_2 \leq C_2\sqrt{q_1 \log \bar{p}_n/n}. \end{aligned}$$

The displays above together with (I.43) imply that there is a positive constant C_8 so that

$$\|\mathbf{e}_j^T(\hat{U}_1 - \Sigma_n\hat{\Gamma}_1\hat{\Phi}_1^T)\|_2 \leq C_8 c_{1n}\sqrt{q_1 \log \bar{p}_n/n} \leq \lambda_1 w_j,$$

which contradicts (I.44). Therefore, $P(\hat{\mathcal{A}}_1 \subseteq \mathcal{A}_1) \to 1$ as $n \to \infty$. We complete the proof of Theorem 4.1.

□

Proof of Theorem 4.3. By the similar argument as (I.34), we obtain that

$$G_1 + G_2 + G_3 + \lambda_1 \sum_{j \in \check{\mathcal{A}}_1^c} \|\hat{\boldsymbol{\delta}}_{1j}\|_2 \leq \lambda_1 \sum_{j \in \check{\mathcal{A}}_1} \|\hat{\boldsymbol{\delta}}_{1j}\|_2 + 2\lambda_1 \sum_{j \in \check{\mathcal{A}}_1^c} \|\boldsymbol{\gamma}_{1j}\|_2.$$

In the following, we assume (I.35) and (I.36) hold. Then the display above implies that

$$\begin{aligned} & G_3 + \sum_{j \in \check{\mathcal{A}}_1^c} \big(\lambda_1 w_j - c_{1n}\sqrt{\frac{\log \bar{p}_n}{n}}\big)\|\hat{\boldsymbol{\delta}}_{1j}\|_2 \\ \text{(I.45)} \qquad & \leq \sum_{j \in \check{\mathcal{A}}_1} \big(\lambda_1 w_j + c_{1n}\sqrt{\frac{\log \bar{p}_n}{n}}\big)\|\hat{\boldsymbol{\delta}}_{1j}\|_2 + 2\lambda_1 \sum_{j \in \check{\mathcal{A}}_1^c} \|\boldsymbol{\gamma}_{1j}\|_2. \end{aligned}$$

With our choice of λ_1, (I.45) implies that

$$\text{(I.46)} \qquad \sum_{j \in \check{\mathcal{A}}_1^c} \|\hat{\boldsymbol{\delta}}_{1j}\|_2 \leq 2 \sum_{j \in \check{\mathcal{A}}_1} \|\hat{\boldsymbol{\delta}}_{1j}\|_2 + 3\|\Gamma_{1,\check{\mathcal{A}}_1^c}\|_{2,1}.$$

First, we assume that $\sum_{j \in \check{\mathcal{A}}_1^c}\|\hat{\boldsymbol{\delta}}_{1j}\|_2 \leq 3\sum_{j \in \check{\mathcal{A}}_1}\|\hat{\boldsymbol{\delta}}_{1j}\|_2$. Then by (4.5), (I.45) and (I.38),

$$G_4 \leq 4c_{1n}\sqrt{\frac{\log \bar{p}_n}{n}} \sum_{j \in \check{\mathcal{A}}_1} \|\hat{\boldsymbol{\delta}}_{1j}\|_2 + 6c_{1n}\theta\sqrt{\frac{\log \bar{p}_n}{n}} + 8C_2\sqrt{\frac{\log \bar{p}_n}{n}}\big(\sum_{j \in \check{\mathcal{A}}_1} \|\hat{\boldsymbol{\delta}}_{1j}\|_2\big)^2.$$

Define $\grave{\mathcal{A}}_1 = \check{\mathcal{A}}_1 \cup \check{\mathcal{A}}_{11}$, where $\check{\mathcal{A}}_{11}$ is the index set in $\check{\mathcal{A}}_1^c$ that corresponds to the q_1 largest $\|\hat{\boldsymbol{\delta}}_{1j}\|_2$ with $j \in \check{\mathcal{A}}_1^c$. Since $G_4 \geq \frac{1}{2}\sigma_* \sum_{j \in \grave{\mathcal{A}}_1}\|\hat{\boldsymbol{\delta}}_{1j}\|_2^2$, the display above implies that with large enough n,

$$\frac{1}{4}\sigma_* \sum_{j \in \grave{\mathcal{A}}_1} \|\hat{\boldsymbol{\delta}}_{1j}\|_2^2 - 4c_{1n}\sqrt{\frac{q_1 \log \bar{p}_n}{n}}\big(\sum_{j \in \grave{\mathcal{A}}_1} \|\hat{\boldsymbol{\delta}}_{1j}\|_2^2\big)^{1/2} - 6c_{1n}\theta\sqrt{\frac{\log \bar{p}_n}{n}}.$$

Consequently, we have by the quadratic formula that

$$\text{(I.47)} \qquad \big(\sum_{j \in \grave{\mathcal{A}}_1} \|\hat{\boldsymbol{\delta}}_{1j}\|_2^2\big)^{1/2} \leq 16(\sigma_*^{-1} + \sigma_*^{-1/2})c_{1n}\sqrt{\frac{q_1 \log \bar{p}_n}{n}} + 2\sigma_*^{-1/2}\theta.$$

Then by (I.47) and the similar argument as (I.40), we have that

$$\text{(I.48)} \quad \|\hat{\Gamma}_1\hat{\Phi}_1 - \Gamma_1\Phi_1\|_F \leq 64(\sigma_*^{-1} + \sigma_*^{-1/2})^2 c_{1n}\sqrt{q_1 \log \bar{p}_n/n} + 9\sigma_*^{-1/2}\theta.$$

Next, we assume that $\sum_{j \in \check{\mathcal{A}}_1^c}\|\hat{\boldsymbol{\delta}}_{1j}\|_2 > 3\sum_{j \in \check{\mathcal{A}}_1}\|\hat{\boldsymbol{\delta}}_{1j}\|_2$. Then by (I.46),

$$\text{(I.49)} \qquad \sum_{j=1}^{p}\|\hat{\boldsymbol{\delta}}_{1j}\|_2 = \sum_{j \in \check{\mathcal{A}}_1} \|\hat{\boldsymbol{\delta}}_{1j}\|_2 + \sum_{j \in \check{\mathcal{A}}_1^c} \|\hat{\boldsymbol{\delta}}_{1j}\|_2 \leq 4\|\Gamma_{1,\check{\mathcal{A}}_1^c}\|_{2,1} \leq 4\theta.$$

In together with the probability bounds for (I.35) and (I.36), by (I.48), (I.49) and (A5), we have that (4.6) holds. The second statement of Theorem 4.3 follows from the same procedures as the proof of Theorem 4.2. We complete the proof of Theorem 4.3. $\square$

I.5. *Additional theorems and proofs on estimating $Q(t \mid \mathbf{x})$.* As mentioned in the end of Section 4, we may use alternative nonparametric estimation methods such as histogram-based local KM to estimate $Q(t \mid \mathbf{x})$. In the following, we obtain estimation error bound in high probability for histogram in Theorem I.3, which can be subsequently used to establish the consistency results like that of Theorem 4.4. Taking notations as in proof of Theorem 4.4, and suppose we partition support $\mathcal{Z}$ into M sub-domains $\mathcal{B}_1, \cdots, \mathcal{B}_M$ with maximal width $2h_n$. For example, we can use the maximal packing set $\mathcal{Q}_\eta = \{\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_M\}$ with $\eta = h_n/4$ as defined in the proof of Theorem 4.4, and under l_2 norm, let $\mathcal{B}_j$ $(j = 1, \cdots, M)$ be the set of points whose closest point in $\mathcal{Q}_\eta$ is $\boldsymbol{\theta}_j$.

THEOREM I.3. *Suppose Conditions (C1)-(C3) are satisfied. Given a basis Γ_0 of $\mathcal{S}_{T|\mathbf{X}}$, if $\|\Gamma_*^T\mathbf{x} - \Gamma_0^T\mathbf{x}\|_F \leq \Delta_n$ for every $\mathbf{x} \in \mathcal{X}$, then there is a generic constant $\tilde{\alpha} > 0$ so that for every $\epsilon_n \geq \tilde{\alpha}\sqrt{\frac{\log n}{nh_n^d}}$ and large enough n, we have that with probability great than $1 - \tilde{c}_{11}h_n^{-d}\exp(-\tilde{c}_{12}nh_n^d) - \tilde{c}_{21}\epsilon_n^{-2}h_n^{-d}\exp(-\tilde{c}_{22}nh_n^d\epsilon_n^2)$,*

$$\sup_{t\in[0,T_0)}\sup_{\mathbf{x}\in\mathcal{X}}|\hat{Q}_*(t \mid \mathbf{x}) - Q(t \mid \mathbf{x})| \leq \tilde{c}_\tau\epsilon_n + 4c_l h_n + c_l\Delta_n, \tag{I.50}$$

where $\tilde{c}_\tau, \tilde{c}_{11}, \tilde{c}_{12}, \tilde{c}_{21}, \tilde{c}_{22}$ are some positive constants (not depending on n).

PROOF OF THEOREM I.3. Recall that $\mathbf{Z} = \Gamma_*^T\mathbf{X}$ and $\tilde{\mathbf{Z}} = \Gamma_0^T\mathbf{X}$. Given $\mathbf{z} \in \mathcal{Z}$, let $\mathcal{M}(\mathbf{z})$ be the subdomain containing $\mathbf{z}$. Given $\mathbf{x} \in \mathcal{X}$ with $\mathbf{z} = \Gamma_*^T\mathbf{x}$, define $\hat{g}_1(t \mid \mathbf{z}) = \hat{Q}_*(t \mid \mathbf{x})$, $g_1(t \mid \mathbf{z}) = P(T > t \mid \mathbf{Z} \in \mathcal{M}(\mathbf{z}))$ and $g_2(t \mid \mathbf{z}) = P(T > t \mid \tilde{\mathbf{Z}} \in \mathcal{M}(\mathbf{z}))$. Then by (C2),

$$\begin{aligned}\sup_{\mathbf{x}\in\mathcal{X}}|\hat{Q}_*(t \mid \mathbf{x}) - Q(t \mid \mathbf{x})| &\leq \sup_{\mathbf{z}\in\mathcal{Z}}\big|\hat{g}_1(t \mid \mathbf{z}) - g_2(t \mid \mathbf{z})\big| + \sup_{\mathbf{z}\in\mathcal{Z}}\big|g_2(t \mid \mathbf{z}) - m(t \mid \mathbf{z})\big| \\ &\leq \sup_{\mathbf{z}\in\mathcal{Z}}\big|\hat{g}_1(t \mid \mathbf{z}) - g_2(t \mid \mathbf{z})\big| + c_l h_n. \end{aligned}\tag{I.51}$$

Also, note that

$$\begin{aligned}\big|\hat{g}_1(t \mid \mathbf{z}) - g_2(t \mid \mathbf{z})\big| &\leq \big|\hat{g}_1(t \mid \mathbf{z}) - g_1(t \mid \mathbf{z})\big| + \big|g_1(t \mid \mathbf{z}) - g_2(t \mid \mathbf{z})\big| \\ &\leq \big|\hat{g}_1(t \mid \mathbf{z}) - g_1(t \mid \mathbf{z})\big| + c_l\Delta_n + c_l h_n, \end{aligned}\tag{I.52}$$

where the last inequality holds because

$$\begin{aligned}\big|g_1(t \mid \mathbf{z}) - g_2(t \mid \mathbf{z})\big| &\leq \big|\mathrm{E}[m_0(t \mid \tilde{\mathbf{Z}}) \mid \mathbf{Z} \in \mathcal{M}(\mathbf{z})] - \mathrm{E}[m_0(t \mid \mathbf{Z}) \mid \mathbf{Z} \in \mathcal{M}(\mathbf{z})]\big| \\ &\quad + \big|\mathrm{E}[m_0(t \mid \mathbf{Z}) \mid \mathbf{Z} \in \mathcal{M}(\mathbf{z})] - \mathrm{E}[m_0(t \mid \tilde{\mathbf{Z}}) \mid \tilde{\mathbf{Z}} \in \mathcal{M}(\mathbf{z})]\big| \\ &\leq \mathrm{E}[c_l\|\tilde{\mathbf{Z}} - \mathbf{Z}\|_2 \mid \mathbf{Z} \in \mathcal{M}(\mathbf{z})] + c_l h_n \\ &\leq c_l\Delta_n + c_l h_n. \end{aligned}$$

Then, by (I.51) and (I.52),

$$\sup_{t\in[0,T_0)}\sup_{\mathbf{x}\in\mathcal{X}}|\hat{Q}_*(t\,|\,\mathbf{x})-Q(t\,|\,\mathbf{x})| \le \sup_{t\in[0,T_0)}\sup_{\mathbf{z}\in\mathcal{Z}}\big|\hat{g}_1(t\,|\,\mathbf{z})-g_1(t\,|\,\mathbf{z})\big|+2c_lh_n+c_l\Delta_n,$$

which, in together with Lemma I.5, implies (I.50). This completes the proof of Theorem I.3. □

LEMMA I.5. *Suppose conditions (C1)-(C3) are satisfied. Using definitions given in Theorem I.3 and its proof, we have that for every* $\epsilon_n \ge \tilde{\alpha}\sqrt{\frac{\log n}{nh_n^d}}$ *and large enough n, with probability greater than* $1-\tilde{c}_{11}h_n^{-d}\exp(-\tilde{c}_{12}nh_n^d)-\tilde{c}_{21}\epsilon_n^{-2}h_n^{-d}\exp(-\tilde{c}_{22}nh_n^d\epsilon_n^2)$,

$$\sup_{t\in[0,T_0)}\sup_{\mathbf{z}\in\mathcal{Z}}\big|\hat{g}_1(t\,|\,\mathbf{z})-g_1(t\,|\,\mathbf{z})\big| \le c_\tau\epsilon_n+2c_lh_n,$$

where $\tilde{c}_\tau,\tilde{c}_{11},\tilde{c}_{12},\tilde{c}_{21},\tilde{c}_{22}$ *are some positive constants (not depending on n).*

PROOF OF LEMMA I.5. For every $j\in\{1,\cdots,M\}$, define that $H_{1j}(t)=P(Y>t,\,\delta=1\,|\,\mathbf{Z}\in\mathcal{B}_j)$, $H_{2j}=P(Y>t\,|\,\mathbf{Z}\in\mathcal{B}_j)$ and $Q_j(t)=P(T>t\,|\,\mathbf{Z}\in\mathcal{B}_j)$. Also define a function similar to cumulative hazard function $\Lambda_j(t):=-\int_0^t\frac{dH_{1j}(s)}{H_{2j}(s)}$. Note that by Condition (C1), for every $t\in[0,T_0)$, $g_1(t\,|\,\mathbf{z})\ge\tau_0$ and $H_{2j}\ge\tau_0^2=:\tau_1$. Define

$$\hat{H}_{1j}(t)=\frac{\sum_{i=1}^n I(Y_i>t,\delta_i=1,\mathbf{Z}_i\in\mathcal{B}_j)}{\sum_{i=1}^n I(\mathbf{Z}_i\in\mathcal{B}_j)},\ \hat{H}_{2j}(t)=\frac{\sum_{i=1}^n I(Y_i>t,\mathbf{Z}_i\in\mathcal{B}_j)}{\sum_{i=1}^n I(\mathbf{Z}_i\in\mathcal{B}_j)},$$

and

$$\hat{\Lambda}_j(t)=-\int_0^t\frac{d\hat{H}_{1j}(s)}{\hat{H}_{2j}(s-)}=\sum_{k=1}^n\frac{I(Y_k\le t,\,\delta_k=1,\,\mathbf{Z}_k\in\mathcal{B}_j)}{\sum_{i=1}^n I(Y_i>Y_k,\,\mathbf{Z}_i\in\mathcal{B}_j)}.$$

Then we can see by definition of $\hat{g}_1(t\,|\,\mathbf{z})$ that for any $\mathbf{z}\in\mathcal{B}_j$,

$$\hat{g}_1(t\,|\,\mathbf{z})=\prod_{k=1}^n\Big\{1-\frac{I(Y_k\le t,\,\delta_k=1,\,\mathbf{Z}_k\in\mathcal{B}_j)}{\sum_{i=1}^n I(Y_i\ge Y_k,\,\mathbf{Z}_i\in\mathcal{B}_j)}\Big\}=\prod_{s\le t}^{n}\{1-d\hat{\Lambda}_j(s)\},$$

where $\prod$ is the product integral (Gill & Johansen 1990). By Proposition A.4.1 in Gill (1980), for every $t\in[0,T_0)$,

$$\frac{\hat{g}_1(t\,|\,\mathbf{z})}{g_1(t\,|\,\mathbf{z})}=1-\int_0^t\frac{\hat{g}_1(s-\,|\,\mathbf{z})}{g_1(s\,|\,\mathbf{z})}d[\hat{\Lambda}_j(s)-\Lambda_j(s)],$$

which implies by Theorem 7.3.1 in Shorack & Wellner (2009) that

$$\begin{aligned}|\hat{g}_1(t\,|\,\mathbf{z})-g_1(t\,|\,\mathbf{z})| &\le \Big|\int_0^t\frac{\hat{g}_1(s-|\mathbf{z})}{g_1(s|\mathbf{z})}d[\hat{\Lambda}_j(s)-\Lambda_j(s)]\Big| \\ &\le \frac{3}{\tau_0}|\hat{\Lambda}_j(s)-\Lambda_j(s)|. \qquad \text{(I.53)}\end{aligned}$$

Also, note that under the event $\mathcal{G}_j := \{\|\hat{H}_{2j} - H_{2j}\|_{T_0} \le \tau_1/2\}$

$$|\hat{\Lambda}_j(s) - \Lambda_j(s)| = \Big|\int_0^t \frac{d\hat{H}_{1j}(s)}{\hat{H}_{2j}(s-)} - \int_0^t \frac{dH_{1j}(s)}{H_{2j}(s)}\Big|$$

$$\le \Big|\int_0^t (\frac{1}{\hat{H}_{2j}(s-)} - \frac{1}{H_{2j}(s)}) d\hat{H}_{1j}(s)\Big| + \Big|\int_0^t \frac{1}{H_{2j}(s)} d[\hat{H}_{1j}(s) - H_{1j}(s)]\Big|$$

(I.54)

$$\le 8\tau_1^{-2}\|\hat{H}_{2j} - H_{2j}\|_{T_0} + 2\tau_1^{-1}\|\hat{H}_{1j} - H_{1j}\|_{T_0}.$$

Then under $\mathcal{G}_j$, the two displays above imply that

(I.55)

$$B_{1j} := \sup_{t\in[0,T_0)} |\hat{g}_1(t\,|\,\mathbf{z}) - g_1(t\,|\,\mathbf{z})| \le \frac{24}{\tau_0\tau_1^2}\|\hat{H}_{2j} - H_{2j}\|_{T_0} + \frac{6}{\tau_0\tau_1}\|\hat{H}_{1j} - H_{1j}\|_{T_0}.$$

Next, we want to provide upper bound for B_{1j}. Let N_j be the number of observations in $\mathcal{B}_j$. By the extended Bernstein inequality, there is a constant $c_u > 0$ such that

$$P\big(N_j \le \frac{nc_u h_n^{d_1}}{2}\big) \le \exp\big(-\frac{3c_u n h_n^{d_1}}{28}\big). \tag{I.56}$$

Suppose $\epsilon \to 0$ and $n^\nu \epsilon_n \to \infty$ for some $\nu \ge 1/2$. Then, note by (I.5) in Lemma I.3 that there is a constant $\tilde{\alpha} > 0$ such that for every $j \ge \frac{nc_u h_n^d}{2}$, if $\epsilon_n \ge \tilde{\alpha}\sqrt{\frac{\log n}{nh_n^d}}$, then $\epsilon_n \ge 2\sqrt{\frac{\log N_{\epsilon_n}}{j}}$. Indeed, for any $\tilde{\alpha} \ge 4\sqrt{\nu a_1/c_u}$ and large enough n,

$$j\epsilon_n^2 \ge \frac{nc_u h_n^d \epsilon_n^2}{2} \ge \frac{\tilde{\alpha}^2 c_u \log n}{2} \ge \frac{\tilde{\alpha}^2 c_u}{2\nu}\log(\frac{1}{\epsilon_n}) \ge \frac{\tilde{\alpha}^2 c_u}{4\nu a_1}\log N_{\epsilon_n} \ge 4\log N_{\epsilon_n}.$$

Then, take $\epsilon_n \ge \alpha\sqrt{\frac{\log n}{nh_n^d}}$ with $\epsilon_n \to 0$. By (I.56) and Lemma I.4, defining $n_1 = \lceil \frac{nc_u h_n^d}{2} \rceil$, we have

$$\begin{aligned} &P(\|\hat{H}_{2j} - H_{2j}\|_{T_0} > 8\epsilon_n) \\ \le &P\big(N_j \le \frac{nc_u h_n^d}{2}\big) + \sum_{k=n_1}^n P(\|\hat{H}_{2j} - H_{2j}\|_{T_0} > 8\epsilon_n,\, N_j = k) \\ \le &\exp\big(-\frac{3c_u n h_n^d}{28}\big) + \sum_{k=n_1}^n 8\exp\big(-\frac{k\epsilon_n^2}{4}\big) \\ \le &\exp\big(-\frac{3c_u n h_n^d}{28}\big) + \frac{64}{\epsilon_n^2}\exp\big(-\frac{n_1\epsilon_n^2}{4}\big) =: \tilde{\kappa}_n. \end{aligned} \tag{I.57}$$

Similarly, we can show that

$$P(\|\hat{H}_{1j} - H_{1j}\|_{T_0} > 8\epsilon_n) \le \tilde{\kappa}_n. \tag{I.58}$$

Consequently, by (I.55), (I.57) and (I.58), defining $\tilde{c}_\tau = \frac{192}{\tau_0\tau_1^2} + \frac{48}{\tau_0\tau_1}$,

$$P(B_{1j} \geq \tilde{c}_\tau\epsilon_n) \leq P(\mathcal{G}_j^c) + P(B_{1j} \geq \tilde{c}_\tau\epsilon_n, \mathcal{G}_j)$$

(I.59)

$$\leq P(\mathcal{G}_j^c) + P(\|\hat{H}_{2j} - H_{2j}\|_{T_0} > 8\epsilon_n) + P(\|\hat{H}_{1j} - H_{1j}\|_{T_0} > 8\epsilon_n) \leq 3\tilde{\kappa}_n.$$

By $M \leq (12K/h_n)^d$, (I.59) and union bound, we conclude that Lemma I.5 holds. □

Next, we present an estimation error bound in high probability for the estimation of $S(t \,|\, \mathbf{x})$ in the following lemma, which is useful for establishing consistency of DASH estimator in Thoerem 4.2. Either histogram or Nadaraya-Watson based KM estimator may be used. Given $\Gamma_* \in \mathbb{R}^{p\times d_1}$, denote $\hat{S}_*(t \,|\, \mathbf{x})$ by setting $\hat{\Gamma}_1 = \Gamma_*$ in (3.7). Suppose Δ_n, h_n and ϵ_n are decreasing sequences with $\Delta_n \to 0$, $h_n \to 0$ and $\epsilon_n \to 0$ as $n \to \infty$. The proof is similar to that of Theorem I.3 and thus omitted.

LEMMA I.6. *Suppose Conditions (C1)–(C3) are satisfied. Given a basis Γ_1 of $\mathcal{S}_{(T,C)|\mathbf{X}}$, if $\|\Gamma_*^T\mathbf{x} - \Gamma_1^T\mathbf{x}\|_F \leq \Delta_n$ for every $\mathbf{x} \in \mathcal{X}$, then there is a constant $\alpha > 0$ such that for every $\epsilon_n \geq \alpha\sqrt{\frac{\log n}{nh_n^{d_1}}}$ and large enough n, given any $\mathbf{x}_0 \in \mathcal{X}$, we have with probability great than $1 - c_{11}\exp(-c_{12}nh_n^{d_1}) - c_{21}\epsilon_n^{-2}\exp(-c_{22}nh_n^{d_1}\epsilon_n^2)$ that,*

$$\sup_{t\in[0,T_0)} |\hat{S}_*(t \,|\, \mathbf{x}_0) - S(t \,|\, \mathbf{x}_0)| \leq c_\tau\epsilon_n + 4c_l h_n + c_l\Delta_n, \tag{I.60}$$

where $c_\tau, c_{11}, c_{12}, c_{21}, c_{22}$ are some generic positive constants.

II. Additional details on computation and numerical studies.

Supplement II.1 describes our iterative algorithm to solve optimization problem for the objective function (3.2). Supplement II.2 gives additional simulation studies under various scenarios, and Supplement II.3 provides pre-processing and additional numerical results for the KIRC data.

II.1. *Iterative algorithm for solving (3.2).*

We describe here the iterative algorithm used to solve the objective function (3.2). Let $(\Gamma_{(t-1)}, \Phi_{(t-1)})$ be the estimator after the $(t-1)$th iteration. To update Γ at the tth iteration, define $F_{0n}(\Gamma, \Phi) = \mathrm{tr}\{(\hat{U}_c - \hat{\Sigma}_n\Gamma\Phi)^T\hat{\Omega}_n(\hat{U}_c - \hat{\Sigma}_n\Gamma\Phi)\}$, $H = \hat{\Sigma}_n\hat{\Omega}_n\hat{\Sigma}_n$ and

$$\begin{aligned} G^{(t-1)}(\Gamma) :=& \frac{\partial F_{0n}(\Gamma, \Phi_{(t-1)})}{\partial \mathrm{vec}(\Gamma^T)} \\ =& -2\{I_p \otimes (\Phi_{(t-1)}\hat{U}_c^T)\}\mathrm{vec}(\hat{\Omega}_n\hat{\Sigma}_n) + 2(\hat{\Sigma}_n \otimes I_d)\mathrm{vec}(\Gamma^T). \end{aligned}$$

Consider a quadratic approximation of $F_n(\Gamma, \hat{\Phi}_{(t-1)})$ as

(II.1)

$$F_n^{(t-1)}(\Gamma) = G_{(t-1)}^T\big\{\mathrm{vec}(\Gamma^T) - \mathrm{vec}(\Gamma_{(t-1)}^T)\big\} + \big\{\mathrm{vec}(\Gamma^T) - \mathrm{vec}(\Gamma_{(t-1)}^T)\big\}^T \bar{H}\big\{\mathrm{vec}(\Gamma^T) - \mathrm{vec}(\Gamma_{(t-1)}^T)\big\} + \lambda \sum_{j=1}^{p} w_j \|\boldsymbol{\gamma}_j\|_2,$$

where $G_{(t-1)} = G^{(t-1)}(\Gamma_{(t-1)})$, $\bar{h} = \lambda_{\max}(H)$ and $\bar{H} = \bar{h} I_{pd}$. Then we find $\Gamma_{(t)}$ by $\Gamma_{(t)} = \arg\min_\Gamma F_n^{(t-1)}(\Gamma)$. In this step, we use the surrogate function $F_n^{(t-1)}(\Gamma)$ for optimization rather than $F_n(\Gamma, \Phi_{(t-1)})$ because of two reasons. First, $F_n^{(t-1)}(\Gamma)$ is a majorized surrogate function (see, e.g., Wu & Lange 2010, Qian et al. 2016) of $F_n(\Gamma, \Phi_{(t-1)})$. Its optimization is known to drive downhill the targeted objective, that is, $F_n(\Gamma_{(t)}, \Phi_{(t-1)}) \leq F_n(\Gamma_{(t-1)}, \Phi_{(t-1)})$, leading to stable numerical performance. Second, the minimizer of (II.1) has a convenient closed form solution and all rows of $\Gamma_{(t)}$ can be updated simultaneously. Indeed, by KKT condition, it is not hard to obtain the jth row of $\Gamma_{(t)}$ as $\boldsymbol{\gamma}_j^{(t)} = \frac{1}{\bar{h}}(1 - \frac{\lambda w_j}{\|\boldsymbol{\psi}_j^{(t-1)}\|_2})_+ \boldsymbol{\psi}_j^{(t-1)}$, where $\boldsymbol{\psi}_j^{(t)} = \Phi_{(t)} \hat{U}_c^T \boldsymbol{\ell}_j - \sum_{k=1}^p h_{jk} \boldsymbol{\gamma}_k^{(t)} + \bar{h}\boldsymbol{\gamma}_j^{(t)}$, $\boldsymbol{\ell}_j = \hat{\Omega}_n \hat{\Sigma}_n \mathbf{e}_j$, $h_{jk} = (H)_{jk}$ and $t_+ = t \vee 0$. To update Φ at the tth iteration, we employ the Stiefel manifold optimization on $F_n(\Gamma_{(t)}, \Phi)$ with a Procrustes rotation to obtain $\Phi_{(t)} = R_{(t)} L_{(t)}^T$, where $L_{(t)} \in \mathbb{R}^{b \times d}$, $R_{(t)} \in \mathbb{R}^{d \times d}$, and $L_{(t)} D_{(t)} R_{(t)}^T$ is the SVD of $\hat{U}_c^T \hat{\Omega}_n \hat{\Sigma}_n \Gamma_{(t)}$. Note that this algorithm depends on $\hat{\Omega}_n$ only through $\hat{\Omega}_n \hat{\Sigma}_n$, and by setting $\hat{\Omega}_n \hat{\Sigma}_n = I_p$ like their population counterparts, the optimization does not require precision matrix estimation.

II.2. *Additional simulation results.*

II.2.1. *Conditional quantile estimation.* Following Section 6.2, we provide simulation results on estimating conditional quantile functions under nonlinear and heteroscedastic scenarios. We simulated survival time T from the following two models: Case 3: $T = \exp(-1.5 + \boldsymbol{\beta}^T \mathbf{X} + 0.5\varepsilon)$ and Case 4: $T = \exp\{-2.5 + \boldsymbol{\beta}^T \mathbf{X} + (0.35 + 0.2\boldsymbol{\beta}^T \mathbf{X})^2 \varepsilon\}$, where the censored time C was generated from the model $C = \exp(-1 + \boldsymbol{\beta}^T \mathbf{X} + 0.5\varepsilon_1)$ for both cases, and all other variables and parameters were simulated in the same way as in Case 1. We used the central subspace $\mathcal{S}_{T|\mathbf{X}} = \mathrm{Span}(\boldsymbol{\beta})$ for convenience of visualization, and the conditional quantile functions become single index models ($d = 1$). We then applied SC-DASH to a randomly simulated dataset. Let $\hat{\Gamma}_0$ denote the estimated basis for $\mathcal{S}_{T|\mathbf{X}}$. We employed the local KM estimator in (3.4) to estimate $Q(T \mid \mathbf{X})$ and correspondingly, the estimated

τ-th conditional quantile $\hat{Q}_T(\tau \mid \mathbf{X})$ for $T \mid \mathbf{X}$. From the simulation setting, we note that the true conditional quantile functions are $Q_T(\tau \mid \mathbf{X}) = \exp\{-1.5 + \boldsymbol{\beta}^T\mathbf{X} + 0.5F^{-1}(\tau)\}$ for Case 3 and $Q_T(\tau \mid \mathbf{X}) = \exp\{-2.5 + \boldsymbol{\beta}^T\mathbf{X} + (0.35 + 0.2\boldsymbol{\beta}^T\mathbf{X})^2 F^{-1}(\tau)\}$ for Case 4, where $F^{-1}(\tau)$ is the τ-th quantile of ε. We can then compare the estimated conditional quantile functions with the true counterparts to evaluate the estimation performance.

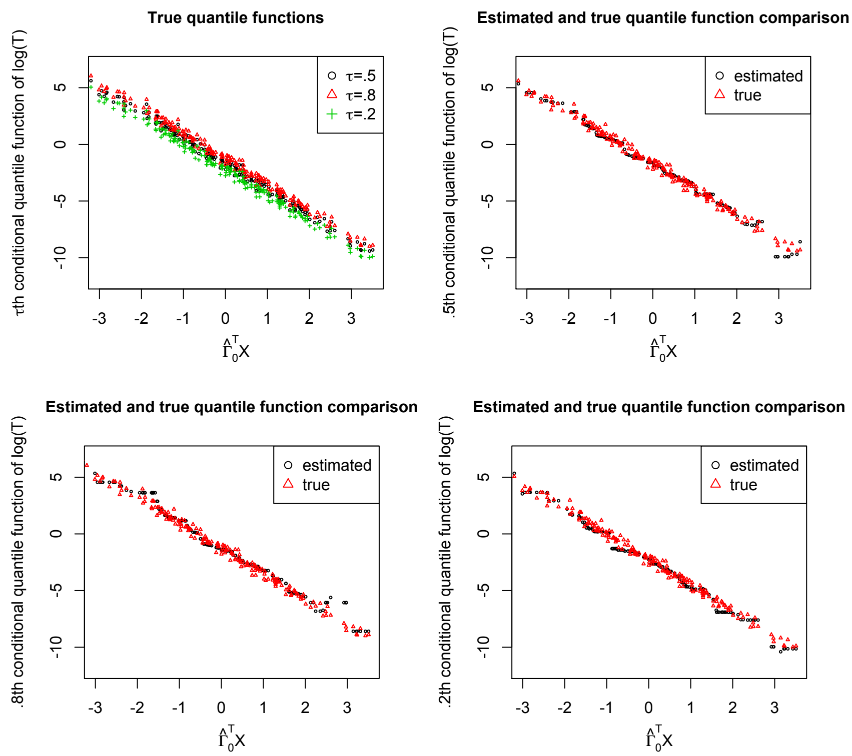

FIG II.1. *Comparison of estimated conditional quantile functions with their true functions at* $\tau = .5, .8, .2$ *for Case 3.*

Figure II.1 and Figure II.2 demonstrate the comparison results over $\tau = 0.5$ (median), 0.8, and 0.2 for Case 3 and Case 4, respectively. In both figures, the top-left subfigures show the true conditional quantile functions of $\log(T_i)$ given $\mathbf{X}_i$ versus $\hat{\Gamma}_0^T\mathbf{X}_i$ at different levels of τ, and the rest of three subfigures compare the true and estimated conditional quantile functions at each level of τ. As expected, the quantile functions of $\log(T)$ given $\mathbf{X}$ in Figure II.1 showed linear patterns for Case 3 while Figure II.2 showed nonlinear patterns for Case 4. Satisfactorily, our method in combination with the local KM estimation (3.4) approximated the true quantile functions reasonably well

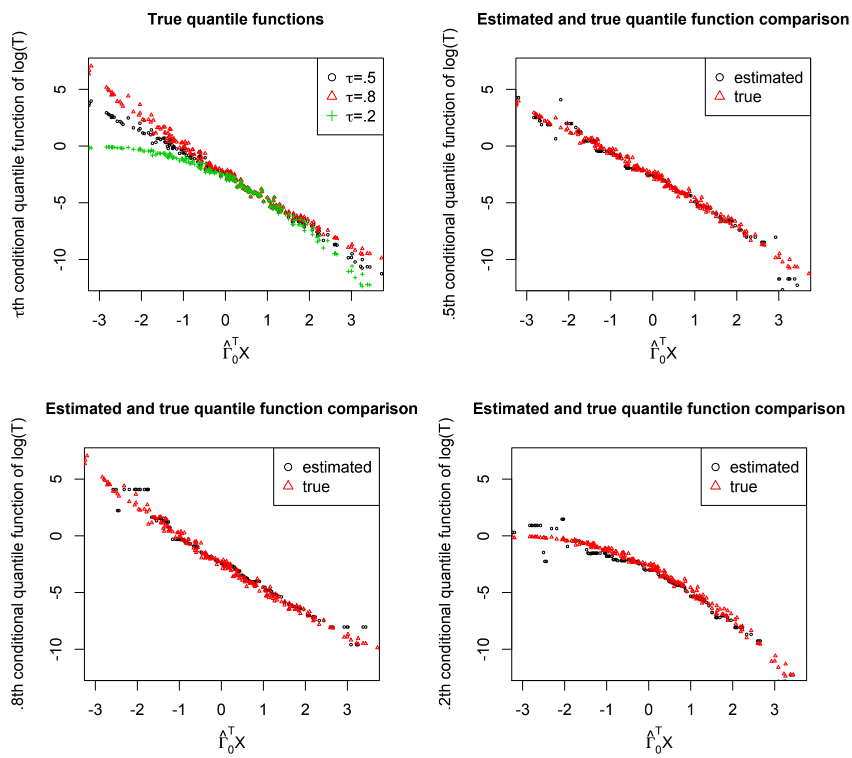

FIG II.2. *Comparison of estimated conditional quantile functions with their true functions at* $\tau = .5, .8, .2$ *for Case 4.*

in both cases.

II.2.2. *Non-continuous predictors.* Following Section 6.3, we consider some simulation situations where the covariates are not necessarily continuous random variables from elliptical distribution family. As one example, we assume the covariates are generated by truncating all elements in $\mathbf{X}$ to be within $[0, 1]$ so that the covariates have probability point mass at -1 and 1. All the other data generating settings remain the same as that of Cases A_0 and Case B_0, and we denote their corresponding scenarios with truncated covariates as Case A_3 and Case B_3, respectively. Satisfactorily, as summarized in Table II.1, despite the slightly larger estimation error than those observed in Cases A_0 and B_0, DASH still performed reasonably well when compared to DS and Coxnet.

In addition, we used Cases 1 and 2 to provide additional numerical results for sensitivity analysis of the linearity condition. Specifically, instead of sim-

Table II.1
Averaged simulation results with non-continuous predictors.

Case	Method	PH			PO		
		Frobenius-norm loss	$\overline{C}_v$	$\overline{IC}_v$	Frobenius-norm loss	$\overline{C}_v$	$\overline{IC}_v$
A_3	Oracle	-	5	0	-	5	0
	Coxnet	1.30 (0.01)	0.86	3.99	1.30 (0.01)	1.00	5.31
	DS	1.28 (0.01)	4.53	34.78	1.30 (0.01)	4.60	32.70
	DASH	0.81 (0.02)	4.46	12.26	0.83 (0.02)	4.55	13.02
B_3	Oracle	-	4	0	-	4	0
	Coxnet	1.30 (0.01)	0.85	5.66	1.31 (0.01)	0.48	2.95
	DS	1.20 (0.01)	3.88	20.51	1.19 (0.01)	3.87	22.09
	DASH	0.63 (0.02)	3.79	4.74	0.54 (0.02)	3.82	4.07

ulating the predictors from multivariate normal distribution, we considered Bernoulli, Binomial, and Poisson distributed predictors. The proposed methods still showed satisfactory results compared to those in original settings, the details are given in Supplement II.2.6 and Table II.4.

II.2.3. *Estimated survival functions and bootstrap intervals.* Following Section 6.3, we observed from Table 2 (and the tables that follow) that DASH performed much better than Coxnet in both estimation and variable selection. To visualize the different performance between DASH and Coxnet here, we used a simulation data set with $n = 200$ from Case A_0 as an example to plot the estimated survival function $\hat{Q}(t \mid \mathbf{x})$ and applied the percentile method to heuristically construct the associated point-wise bootstrap confidence intervals (CI). For each of 100 bootstrap samples, we applied both DASH and Coxnet methods to provide a basis estimation of $\mathcal{S}_{T|\mathbf{X}}$ and the estimated sufficient predictors, and then employed the local KM method to generate the bootstrap estimation for $Q(t \mid \mathbf{x})$. Figures II.4 and II.5 contain the plots for DASH and Coxnet, respectively, and the choice of covariates (denoted by $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_3$) in these plots was the same as described at the beginning of Section 6.3. In all the figures, plot (a) gives the estimated survival function $\hat{Q}(t \mid \mathbf{x})$ given the three different covariates, and plots (b)-(d) give the true survival function $Q(t|\mathbf{x})$, its estimation $\hat{Q}(t|\mathbf{x})$, and the point-wise 90% bootstrap CIs for each of the covariates. In addition, we plotted the mean of $\hat{Q}(t \mid \mathbf{x})$ from 100 simulated data sets in Figure II.3(a) and Figure II.3(b) using DASH and Coxnet, respectively.

Comparing Figures II.4(a) and II.3(a) with Figures II.5(a) and II.3(b), the three estimated survival curves from Coxnet are very close to each other, which reflects the unsatisfactory variable selection results (as seen in Table 2), while DASH appears to better reveal the heterogeneous population

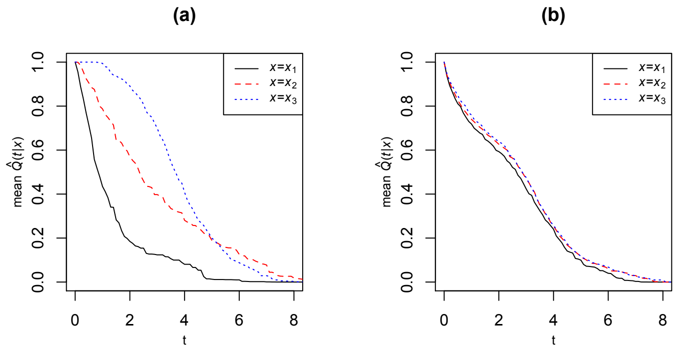

FIG II.3. *Mean of $\hat{Q}(t|\mathbf{x})$ from 100 simulated data sets in Case A. (a): DASH; (b): Coxnet*

as expected from Figure 1. In addition, in Figure II.4(b)-(d), the bootstrap intervals from DASH can largely cover the true survival curves, but this was much less evident for those from Coxnet in Figure II.5(c). We also repeated the experiment above with larger sample size $n = 600$, resulting in narrower bootstrap CIs in Figures II.6 and II.7; as opposed to DASH, much of the true survival curve in Figure II.7(c) remained to fall outside the Coxnet bootstrap CIs. In addition, we made plots for the pointwise 5th percentile and 95th percentile of $\hat{Q}(t|\mathbf{x})$ over 100 simulated data sets, which showed very similar patterns as the bootstrap plots above and is thus omitted.

II.2.4. *Weight updating scheme.* Following Section 6.3, we also consider two possible weight re-estimating schemes after obtaining the DASH estimator $\hat{\Gamma}_0$. In the first scheme, we use $\hat{\Gamma}_0$ to re-compute the penalty weights w_j, and we denote this procedure by DASH-wt1; in the second scheme, we use $\hat{\Gamma}_0$ to repeat the inverse probability weighting (IPCW in Step 2) in Algorithm 1, and we denote this procedure by DASH-wt2. For DASH-wt1, we do not expect a change in estimation convergence rate and our experience is that DASH-wt1 tends to give similar or slightly better results than original DASH. On the other hand, DASH-wt2 may not be desirable when $\mathcal{S}_{T|\mathbf{X}}$ is a proper subset of $\mathcal{S}_{(T,C)|\mathbf{X}}$, and the use of DASH estimator (as opposed to DS) may result in reduced accuracy for estimation of $S(t\,|\,\mathbf{X})$ as required in IPCW weighting. Indeed, we applied DASH-wt1 and DASH-wt2 to Cases A_0 and B_0, and observed from the summarized results in Table II.2 that the estimation performance of DASH-wt1 is slightly better (but not significantly) than that of original DASH; On the other hand, the original DASH can perform better than DASH-wt2 (sometimes significantly as in Case B_0). Therefore, we do not find these weight re-estimating schemes to update

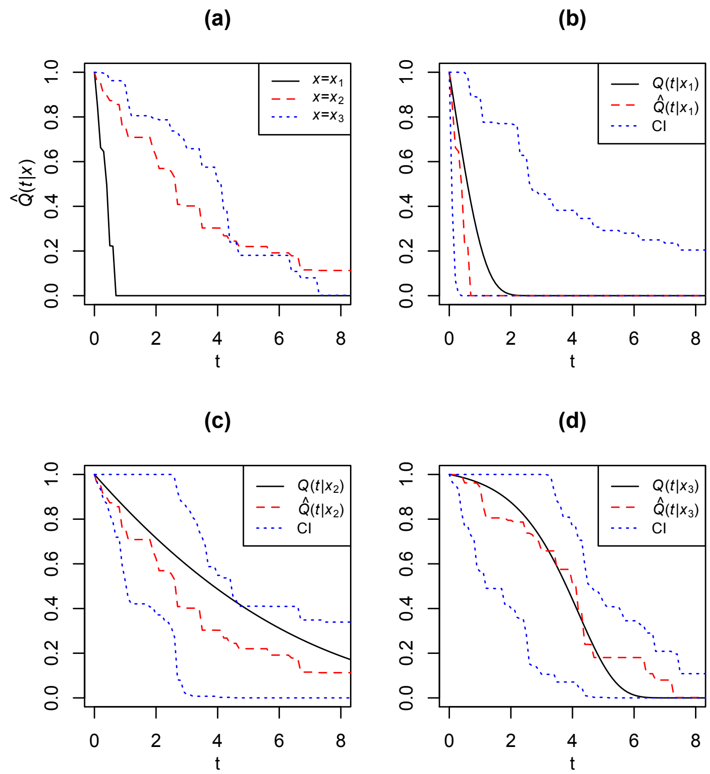

FIG II.4. *Estimated survival functions $\hat{Q}(t|\mathbf{x})$ and empirical pointwise 90% bootstrap confidence intervals (CI) by DASH from a simulated Case A data set with $n = 200$. Plot (a) gives the estimated survival functions $\hat{Q}(t|\mathbf{x})$ for three different covariate values. Plots (b)-(d) gives $Q(t|\mathbf{x})$, $\hat{Q}(t|\mathbf{x})$ and CI for each of the covariates.*

DASH to provide a guaranteed improvement over the original DASH.

TABLE II.2
Averaged simulation results with different weight updating schemes.

Case	Method	PH			PO		
		Frobenius-norm loss	$\overline{C}_v$	$\overline{IC}_v$	Frobenius-norm loss	$\overline{C}_v$	$\overline{IC}_v$
A_0	DASH	0.61 (0.02)	4.84	8.62	0.62 (0.02)	4.76	8.42
	DASH-wt1	0.61 (0.02)	4.84	7.39	0.61 (0.02)	4.76	7.08
	DASH-wt2	0.62 (0.02)	4.83	6.71	0.64 (0.02)	4.74	7.32
B_0	DASH	0.55 (0.02)	3.89	3.65	0.45 (0.02)	3.90	3.17
	DASH-wt1	0.54 (0.02)	3.89	3.00	0.44 (0.02)	3.90	2.60
	DASH-wt2	0.75 (0.02)	3.79	3.08	0.53 (0.02)	3.87	2.66

Next, we provide some additional simulation results with possibly mis-

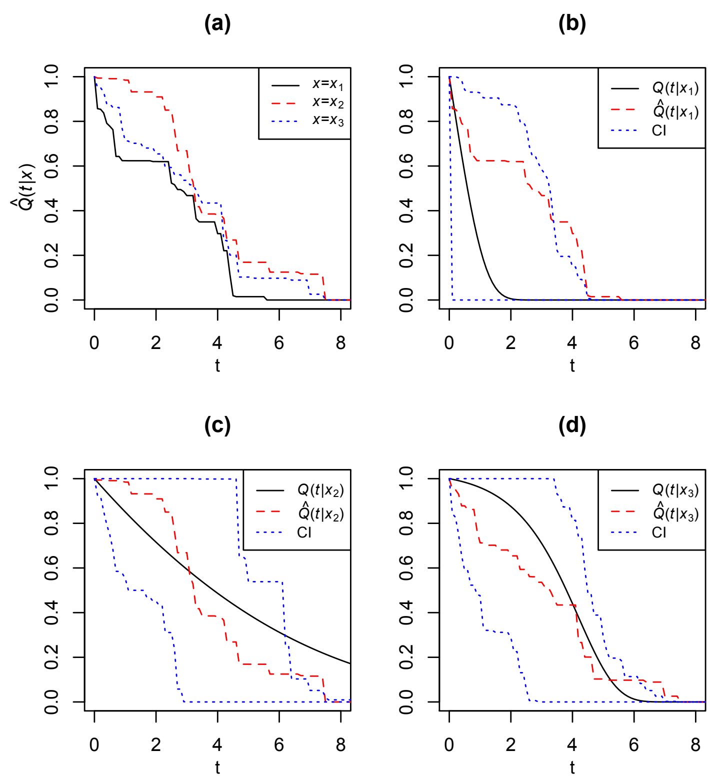

FIG II.5. *Estimated survival functions* $\hat{Q}(t|\mathbf{x})$ *and empirical pointwise* 90% *bootstrap confidence intervals (CI) by Coxnet from a simulated Case A data set with* $n = 200$.

specified structural dimension and discrete covariate scenarios.

II.2.5. *Additional cases with misspecified* d_1. To provide more evaluation of DASH with the under-estimated $dim(\mathcal{S}_{(T,C)|\mathbf{X}})$, we performed additional simulation studies where the true structural dimension of $\mathcal{S}_{(T,C)|\mathbf{X}}$ is larger than 2 but is misspecified (under-estimated) in the estimation procedure. We consider the following settings (Cases 5 and 6) where $p = 1000$ and $n = 200$ and the covariate vector $\mathbf{X}$ was generated in the same way as in Case 1.

Case 5: $T = \exp(-2 + \boldsymbol{\beta}_1^T\mathbf{X} + 0.3\varepsilon)$, $C = \exp\{\text{sign}(\boldsymbol{\beta}_2^T\mathbf{X}) \cdot \log|\boldsymbol{\beta}_3^T\mathbf{X} + 5| + 0.2\varepsilon_1\}$, where $\boldsymbol{\beta}_1 = (1,1,1,1,1,0,\ldots,0)^T$, $\boldsymbol{\beta}_2 = (0,\ldots,0,1,1)^T$, $\boldsymbol{\beta}_3 = (1,1,0,\ldots,0)^T$, and the rest settings are the same as in Case 1. In this example, $\dim(\mathcal{S}_{(T,C)|\mathbf{X}}) = 3$ and $\dim(\mathcal{S}_{T|\mathbf{X}}) = 1$.

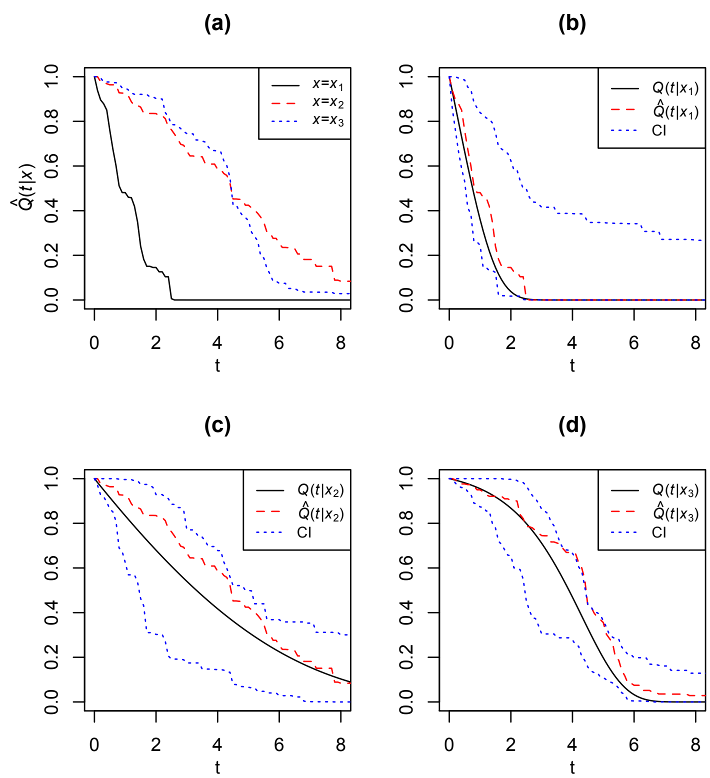

FIG II.6. *Estimated survival functions $\hat{Q}(t|\mathbf{x})$ and empirical pointwise 90% bootstrap confidence intervals (CI) by DASH from a simulated Case A data set with $n = 600$.*

Case 6: $T = \exp\{\text{sign}(\boldsymbol{\beta}_1^T \mathbf{X}) \cdot \log|\boldsymbol{\beta}_2^T \mathbf{X} + 5| + 0.2\varepsilon\}$, $C = \exp\{2.5 + \text{sign}(\boldsymbol{\beta}_3^T \mathbf{X}) \cdot |\boldsymbol{\beta}_4^T \mathbf{X} + 1| + 0.2\varepsilon_1\}$, where $\boldsymbol{\beta}_1 = (-1, 1, 1, -1, 0, \ldots, 0)^T$, $\boldsymbol{\beta}_2 = (1, 1, 1, 1, 0, \ldots, 0)^T$, $\boldsymbol{\beta}_4 = (0, \ldots, 0, 1, 1, 0, \ldots, 0)^T$ with the 989th and 990th elements to be one, and the rest settings are the same as in Case 1. In this example, $\dim(\mathcal{S}_{(T,C)|\mathbf{X}}) = 4$ and $\dim(\mathcal{S}_{T|\mathbf{X}}) = 2$.

We misspecified $d_1 = 2$ for Cases 5 and $d_1 = 3$ for Cases 6 in the estimation, and assume d is known. Table II.3 summarizes the averaged results for these underestimated structural dimension scenarios, which indicates that the DASH estimator can often perform well when the initial DS step has underestimated $\dim(\mathcal{S}_{(T,C)|\mathbf{X}})$.

II.2.6. *Additional cases with non-continuous predictors.* Here, we use Cases 1 and 2 to provide additional numerical results for sensitivity analysis of the linearity condition. Specifically, instead of simulating $\mathbf{X}$ from multivariate normal distribution, we considered the following predictor generation

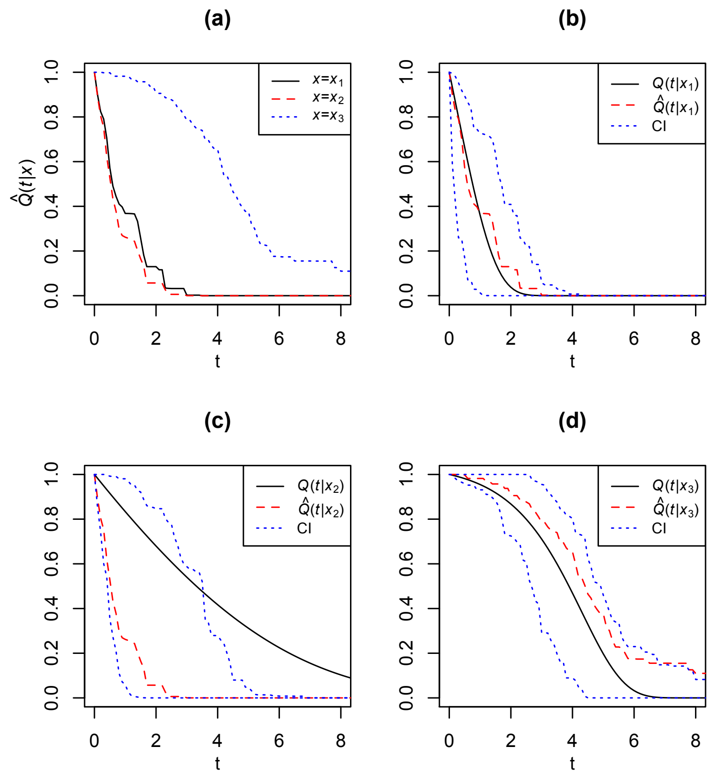

FIG II.7. *Estimated survival functions* $\hat{Q}(t|\mathbf{x})$ *and empirical pointwise* 90% *bootstrap confidence intervals (CI) by Coxnet from a simulated Case A data set with* $n = 600$.

scenarios for Cases 1 and 2. 1) Binary predictors: generated the elements of $\mathbf{X}$ from i.i.d. Bernoulli(.5); 2) Binomial predictors: generated the elements of $\mathbf{X}$ from i.i.d. Binomial(10, .5); 3) Poisson distributed predictors: generated the elements of $\mathbf{X}$ from i.i.d. Poisson(8). In all scenarios, $p = 1000$ and $n = 200$, and the rest settings remain the same as in Cases 1 and 2. Due to the discrete covariates, we used sample covariance estimation only. Table II.4 summarizes the estimation and variable selection results of the three scenarios based on 100 data replications.

The proposed DS and DASH methods perform stably in the three discrete predictor scenarios and the variable selection results remain satisfactory compared to those in original settings, indicating the robustness of our approach to non-continuous predictors in high dimensional settings.

For model misspecification, we also note that the recent work of Chen et al. (2015) proposed a model diagnostic tool for SDR under fixed p large n scenarios with complete response. To our knowledge, model diagnostic for

Table II.3
Averaged simulation results of Cases 5 and 6 based on underestimated dim$(\mathcal{S}_{(T,C)|\mathbf{x}})$ *over 100 runs.*

Cases	Method	Averaged Frobenius-norm loss	Angle between est. and true CS	$\overline{C}_v$	$\overline{IC}_v$
Case 5	Oracle	-	-	5	0
	SC-DS	1.060 (0.004)	13.76	5.00	24.89
	SC-DASH	0.491 (0.021)	20.64	4.86	1.23
	TC-DS	1.064 (0.006)	13.51	4.99	6.52
	TC-DASH	0.438 (0.019)	18.29	4.96	0.91
Case 6	Oracle	-	-	4	0
	SC-DS	1.319 (0.019)	31.23	3.96	24.66
	SC-DASH	0.568 (0.031)	23.05	3.92	3.64
	TC-DS	1.349 (0.027)	29.97	3.98	16.93
	TC-DASH	0.512 (0.027)	19.88	3.96	1.92

SDR with censored data has not been developed in literature; it is not yet clear if any existing tool can be extended to high-dimensional SDR framework (even with complete response). The post selection inference on SDR (with censored data) is also an challenging topic yet to be studied. Though out of scope of the current paper, they are interesting topics that worth investigation in future studies. In addition, the proposed DASH method can be extended to data with matrix- or tensor-valued predictors (e.g., Li et al. 2010, Ding & Cook 2014, 2015*a*,*b*) and integrative multi-source data settings (Jain et al. 2019) which will be studied in future works.

II.3. *Real data pre-processing and additional numerical results.* We performed the KIRC data pre-processing in following steps. For patients with more than one RNA-Seq values, we only kept their first record and excluded low-expression genes with maximum counts of 10. The raw RNA-Seq count data was normalized by the DESeq (Anders & Huber 2010), followed by a logarithmic transformation. Then, for each gene, we truncated the values at three times of its interquartile range above the median. For evaluation of our proposed method, we randomly partition the pre-processed data into a training and testing set with equal size. Using the training set, we further applied univariate gene filtering method by keeping genes with p-values less than 0.001 (Meng et al. 2014) and standardized each gene to have mean 0 and standard deviation 1.

With training set only, the GBM constructed the link function by an ensemble of tree base learners from the sufficient preditors, with the tuning

Table II.4
Results with different types of predictors based on 100 runs.

Cases	Predictor type	Method	Averaged Frobenius-norm loss	Angle between est. and true CS	$\overline{C}_v$	$\overline{IC}_v$
Case 1		Oracle	-	-	5	0
	Binary	DS	0.179 (0.007)	7.28	5.00	0.11
		DASH	0.192 (0.008)	7.83	5.00	0.07
	Binomial	DS	0.129 (0.006)	5.23	5.00	2.21
		DASH	0.182 (0.008)	7.41	5.00	0.58
	Poisson	DS	0.123 (0.006)	5.01	5.00	1.79
		DASH	0.180 (0.007)	7.30	5.00	0.49
Case 2		Oracle	-	-	5	0
	Binary	DS	1.031 (0.003)	9.38	4.99	11.37
		DASH	0.221 (0.011)	9.01	4.99	0.91
	Binomial	DS	1.061 (0.004)	13.70	5.00	28.88
		DASH	0.383 (0.028)	16.40	4.94	2.82
	Poisson	DS	1.052 (0.008)	12.58	4.95	27.43
		DASH	0.296 (0.020)	12.55	4.92	1.73

parameters selected by procedures similar to Yang et al. (2017). We also evaluated the corresponding prediction mean square errors (MSE) defined as follows. Given a time point t, define $\delta_t^\star = I(T > t)$, and let $\hat{Q}(t \mid \mathbf{X})$ be the estimated survival function of $T \mid \mathbf{X}$ using the training set. Define the prediction MSE $g(t) = \mathrm{E}[\delta_t^\star - \hat{Q}(t \mid \mathbf{X})]^2$, which can be estimated based on the testing set to gauge the prediction performance. We used the sample 1st, 2nd,$\cdots$, 99th percentiles of Y_i's as the candidate values of time point t, and computed the relative MSE $g(t) - g_0(t)$, where $g_0(t)$ is the prediction MSE from Coxnet. The boxplots of the relative MSEs are given in Figure II.8. Like AUC, the relative MSEs also showed that DS and DASH performed satisfactorily compared to the benchmark.

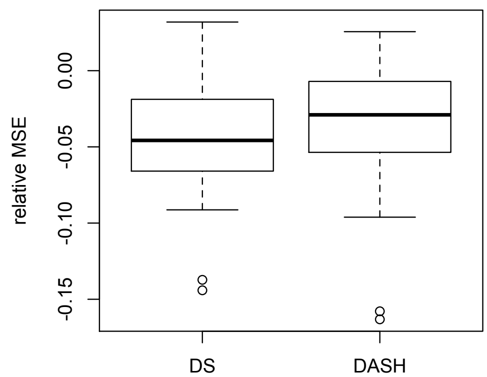

FIG II.8. *Boxplots of the relative MSEs.*

REFERENCES.

Anders, S. & Huber, W. (2010), 'Differential expression analysis for sequence count data', *Genome Biology* **11**(10), R106.

Bickel, P. J. & Levina, E. (2008), 'Covariance regularization by thresholding', *The Annals of Statistics* **6**, 2577–2604.

Chen, X., Cook, R. D. & Zou, C. (2015), 'Diagnostic studies in sufficient dimension reduction', *Biometrika* **102**(3), 545–558.

Dabrowska, D. M. (1989), 'Uniform consistency of the kernel conditional Kaplan-Meier estimate', *The Annals of Statistics* **17**(3), 1157–1167.

Ding, S. & Cook, R. D. (2014), 'Dimension folding PCA and PFC for matrix-valued predictors', *Statistica Sinica* **24**, 463–492.

Ding, S. & Cook, R. D. (2015*a*), 'Higher-order sliced inverse regressions', *Wiley Interdisciplinary Reviews: Computational Statistics* **7**(4), 249–257.

Ding, S. & Cook, R. D. (2015*b*), 'Tensor sliced inverse regression', *Journal of Multivariate Analysis* **133**, 216–231.

Gill, R. D. (1980), 'Censoring and stochastic integrals', *Statistica Neerlandica* **34**(2), 124–124.

Gill, R. D. & Johansen, S. (1990), 'A survey of product-integration with a view toward application in survival analysis', *The Annals of Statistics* **18**(4), 1501–1555.

Jain, Y., Ding, S. & Qiu, J. (2019), 'Sliced inverse regression for integrative multi-omics data analysis', *Statistical Applications in Genetics and Molecular Biology* **18**(1).

Li, B., Kim, M. K. & Altman, N. (2010), 'On dimension folding of matrix-or array-valued statistical objects', *The Annals of Statistics* **38**, 1094–1121.

Meng, J., Li, P., Zhang, Q., Yang, Z. & Fu, S. (2014), 'A four-long non-coding RNA signature in predicting breast cancer survival', *Journal of Experimental & Clinical Cancer Research* **33**(1), 84.

Qian, W. & Yang, Y. (2016), 'Kernel estimation and model combination in a bandit problem with covariates', *Journal of Machine Learning Research* **17**(149), 1–37.

Qian, W., Yang, Y. & Zou, H. (2016), 'Tweedie's compound Poisson model with grouped elastic net', *Journal of Computational and Graphical Statistics* **25**(2), 606–625.

Shorack, G. R. & Wellner, J. A. (2009), *Empirical Processes with Applications to Statistics*, SIAM.

van de Geer, S. (2006), *Empirical Process Theory and Applications*, ETH Zürich.

van der Vaart, A. W. (2000), *Asymptotic Statistics*, Vol. 3, Cambridge University press.

Wu, T. T. & Lange, K. (2010), 'The MM alternative to EM', *Statistical Science* **25**(4), 492–505.

Yang, Y. & Barron, A. (1999), 'Information-theoretic determination of minimax rates of convergence', *Annals of Statistics* **27**, 1564–1599.

Yang, Y., Qian, W. & Zou, H. (2017), 'Insurance premium prediction via gradient tree-boosted Tweedie compound Poisson models', *Journal of Business & Economic Statistics, accepted* .

Yin, X. & Hilafu, H. (2015), 'Sequential sufficient dimension reduction for large p, small n problems', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **77**(4), 879–892.

Yu, Y., Wang, T. & Samworth, R. J. (2015), 'A useful variant of the Davis–Kahan theorem for statisticians', *Biometrika* **102**(2), 315–323.